109 年度初級巨量資料分析師能力鑑定試題
（含疑義題釋覆結果）

科目 2：資料處理與分析概論　　疑義題釋覆結果日期:109.06.16

考試日期：109 年 5 月 30 日　　

單選題 50 題（佔 100%）

D	1. 下列何者「不是」文本語料庫內的文字常見問題？ (A) 格式不良 (B) 非標準化 (C) 不能以二維表格呈現 (D) 為結構化數據
A	2. 一般來說，R 語言通常將遺缺值記為下列何者？ (A) NA（Not Available） (B) NaN（Not a Number） (C) NULL (D) FALSE
B	3. 一般來說，Python 語言通常將遺缺值記為下列何者？ (A) NA（Not Available） (B) NaN（Not a Number） (C) NULL (D) FALSE
D	4. 許多資料匯入軟體環境後，經常需要將資料轉換為特定分析與繪圖方法所需的長格式或是寬格式，這項作業稱為下列何者？ (A) 資料彙總（data aggregation） (B) 資料縮減（data reduction） (C) 資料摘要（data summarization） (D) 資料變形（data reshaping）
B	5. 下列為自然語言處理的基本步驟，下列何者為正確排序？1：斷詞、2：詞性標註、3：專有名詞提取、4：詞組標記、5：斷句 (A) 2＞5＞3＞1＞4 (B) 5＞1＞2＞3＞4 (C) 5＞1＞2＞4＞3 (D) 2＞5＞3＞4＞1
D	6. R 語言中，下列何者「不是」處理群組與摘要的函數？ (A) tapply() (B) summaryBy() (C) aggregate() (D) cut()
D	7. 下列何者「不屬於」非監督式學習？ (A) 關聯法則

	(B) K-Means (C) Word2Vec (D) K Nearest Neighbor
C	8. 參考附圖 R 語言 CO2 資料集，依照 uptake 欄位由大至小排序，下列何者為正確？ Plant Type Treatment conc uptake Qn3 Quebec nonchilled 1000 45.5 Qn2 Quebec nonchilled 1000 44.3 Qn3 Quebec nonchilled 675 43.9 Qn3 Quebec nonchilled 500 42.9 Qc2 Quebec chilled 1000 42.4 Qn3 Quebec nonchilled 350 42.1 (A) order(CO2$uptake) (B) CO2[sort(CO2$uptake),] (C) CO2[order(CO2$uptake, decreasing = TRUE),] (D) CO2[sort(CO2$uptake, decreasing = TRUE),]
D	9. 參考附圖，關於 R 語言資料摘要，下列敘述何者為正確？ > str(mydata) 'data.frame': 1000 obs. of 5 variables: $ SiteId : int 84 83 80 78 77 75 72 71 70 69 ... $ SiteName : Factor w/ 77 levels "二林","三重",..: 51 49 76 44 34 45 42 52 19 $ MonitorDate: Factor w/ 13 levels "2019-12-29","2019-12-30",..: 13 13 13 13 13 $ AQI : int 41 64 28 51 87 71 87 110 53 93 ... $ COSubIndex : int 2 NA NA 3 6 5 9 14 16 7 ... > summary(mydata) SiteId SiteName MonitorDate AQI COSubIndex Min. : 1.00 二林 : 13 2019-12-30: 77 Min. : 14.00 Min. : 1.000 1st Qu.:20.00 三重 : 13 2019-12-31: 77 1st Qu.: 39.00 1st Qu.: 5.000 Median :39.00 三義 : 13 2020-01-01: 77 Median : 56.00 Median : 7.000 Mean :39.65 土城 : 13 2020-01-02: 77 Mean : 63.79 Mean : 7.248 3rd Qu.:59.00 士林 : 13 2020-01-03: 77 3rd Qu.: 85.25 3rd Qu.: 9.000 Max. :84.00 大同 : 13 2020-01-04: 77 Max. :151.00 Max. :26.000 (Other):922 (Other) :538 NA's :78 (A) AQI 欄位的中位數為 63.79 (B) class(mydata$MonitorDate)的結果為"Date" (C) SiteId 欄位的最大值為 59.00 (D) COSubIndex 欄位有 78 個 NA 值
A	10. 若用盒鬚圖（Box plot）來檢視資料時，無法從中觀察到下列何者訊息？ (A) 平均值 (B) 第一四分位數 (C) 最小值 (D) 離群值（Outlier）

109 年度初級巨量資料分析師能力鑑定試題
（含疑義題釋覆結果）

科目 2：資料處理與分析概論　　疑義題釋覆結果日期:109.06.16

考試日期：109 年 5 月 30 日　　

<table>
<tr><td>A</td><td>11. 在機器學習中有多種降低資料維度的方法，下列何者屬於降維度的方法？
(A) 主成份分析（Principal Component Analysis）
(B) 決策樹（Decision Tree）
(C) K-近鄰演算法（K Nearest Neighbor）
(D) 羅吉斯迴歸（Logistic Regression）</td></tr>
<tr><td>C</td><td>12. 下列何者可對連續變量進行離散化（Discretization）處理？
(A) 單熱編碼（One-Hot Encoding）
(B) 標準化（Standardization）
(C) 資料分箱（Binning）
(D) 正規化（Normalization）</td></tr>
<tr><td>D</td><td>13. 參考附圖，是使用下列何者編碼方式對類別資料進行轉換？
<table><tr><th>商品</th><th>價格</th></tr><tr><td>A</td><td>29</td></tr><tr><td>B</td><td>24</td></tr><tr><td>C</td><td>32</td></tr></table>⇨<table><tr><th>A</th><th>B</th><th>C</th><th>價格</th></tr><tr><td>1</td><td>0</td><td>0</td><td>29</td></tr><tr><td>0</td><td>1</td><td>0</td><td>24</td></tr><tr><td>0</td><td>0</td><td>1</td><td>32</td></tr></table>(A) 頻率編碼（Frequency Encoding）
(B) 序號編碼（Ordinal Encoding）
(C) 標籤編碼（Label Encoding）
(D) 單熱編碼（One-Hot Encoding）</td></tr>
<tr><td>A</td><td>14. 關於 Box-Cox 轉換，下列敘述何者正確？
(A) 適用於當變數值恆正的時候
(B) 是一種線性轉換
(C) 可將對稱分佈的變數轉為偏斜分佈
(D) 只適用於右偏的變數分布</td></tr>
<tr><td>C</td><td>15. 附圖是移除預測變數流程中的步驟，下列何者為正確的排序？
1：計算 A 與其他變數間的相關係數平均值，B 亦同
2：如果 A 有較大的平均相關係數，則刪除之。否則，請刪除 B 變數
3：找出相關係數絕對值最大的兩個預測變數 A 與 B
4：重複上述三個步驟，直到沒有相關係數的絕對值超出門檻
5：計算預測變數的相關係數矩陣，並設定相關係數的絕對值門檻</td></tr>
</table>

	(A) 1＞2＞3＞4＞5 (B) 5＞1＞2＞3＞4 (C) 5＞3＞1＞2＞4 (D) 1＞5＞3＞4＞2
D	16. 關於 Hadoop 相關巨量資料處理技術，下列敘述何者「不正確」？ (A) HDFS 是分散式檔案處理架構 (B) HBase 是欄位導向資料模型處理架構 (C) MapReduce 是大量檔案資料處理引擎架構 (D) Hive 是類似 SQL 處理語法以擷取、轉換資料作業
D	17. 關於巨量資料的特性，下列何者「不正確」？ (A) Volume (B) Velocity (C) Variety (D) Visualization
B	18. 相對於單一機器平行運算（Parallel Computing），下列敘述何者「不是」叢集分散式運算（Distributed Computing）的特性？ (A) 可靠性較高 (B) 共享記憶體 (C) 計算效率高 (D) 具可擴展性
D	19. 關於巨量資料處理，下列敘述何者「不正確」？ (A) 分散式架構處理可以提升巨量資料處理效能 (B) 透過 API 或者網路爬蟲的方式，可以來搜集大量外部資料，例如網站資料 (C) 可以根據運算需求與時效性，平行擴增所需要的運算資源，提供更好的運算服務 (D) 許多資料要能即時得到結果才能發揮最大的價值，即為巨量資料所談的「Value」
C	20. 當 Client 端上傳檔案到 HDFS 時，下列敘述何者較為正確？ (A) 資料經過 NameNode 傳給 DataNode (B) 資料區塊將依預先設定依次傳遞 (C) Client 將資料上傳到一台 DataNode 上，並由 DataNode 完成副本的複製工作 (D) 該 DataNode 失敗時，Client 不會繼續上傳給其他 DataNode
C	21. R 語言中，下列何者為計算平均的函數？

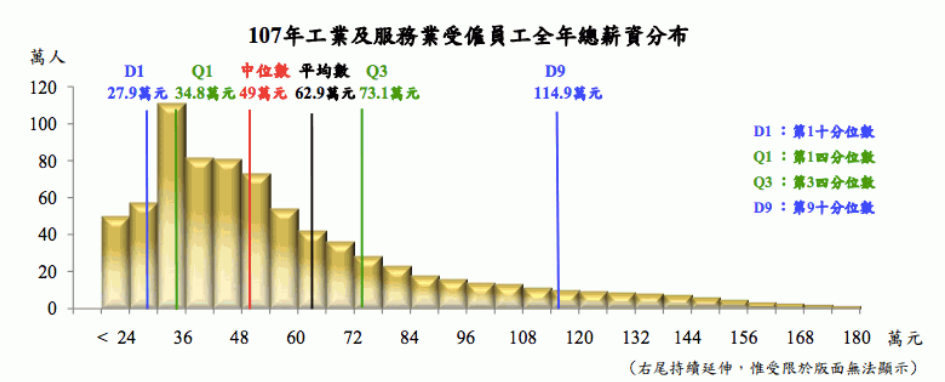
	(A) var() (B) sd() (C) mean() (D) quantile()
A	22. R 語言中，下列何者為計算變異數的函數？ (A) var() (B) sd() (C) mean() (D) quantile()
B	23. 下列何者能夠同時回傳資料向量的最大值與最小值？ (A) diff() (B) range() (C) gather() (D) dcast()
B	24. 要了解預測變數之間的相關程度，可以運用 Python 中 pandas 套件下何者函數？ (A) sum() (B) corr() (C) anova() (D) lm()
D	25. 下列何者屬於連續型變數資料？ (A) 身分證號 (B) 生日 (C) 血型 (D) 體重
D	26. 關於重抽樣方法（resampling methods），下列敘述何者「不正確」？ (A) 重抽樣是反覆地從訓練集或資料集中抽出或有不同的各組樣本，並重新配適各組樣本的模型，以獲得模型相關的額外資訊 (B) 常用的重抽樣方法有拔靴抽樣法（bootstrapping）與 k 摺交叉驗證（k-fold cross validation） (C) 交叉驗證（cross validation）與拔靴抽樣（bootstrapping）兩種方法的差別只在於樣本子集如何被挑出 (D) 一般而言 k 摺交叉驗證（k-fold cross validation）相較於他法有較低的變異，但當訓練集大時則此問題較不嚴重
A	27. 關於相關（correlation）與獨立（independence），下列敘述何者「不

	正確」？ (A) 共變異數是絕對指標，能表達兩變數間關係的強度 (B) 共變異數是最常見的一致性摘要統計量數，它衡量兩變數如何一起變動，亦即同向變動或是反向變動 (C) 將共變異數除以兩隨機變數或變量樣本的標準差即為相關係數 (D) 不一致性指一變量值高（低）時，另一變量反而低（高）
B	28. 下列敘述何者「不正確」？ (A) 理論上常態分配的平均數=中位數 (B) 機率分配中所有事件機率的總和=期望值 (C) p=0.5 的二項分配為對稱型 (D) 卜瓦松分配中事件發生的最小次數為 0
C	29. 『薪資』資料集中的觀察值（單位千元），依遞增順序顯示為：28, 30, 32, 35, 35, 40, 45, 47, 47, 80。下列何者為上述資料中之中位數？ (A) 35 (B) 40 (C) 37.5 (D) 42.5
~~B~~ BC 均給分	30. 關於類別型變數的頻繁次數比（Frequent ratio）與唯一值百分比（Percent unique）資料相關性，下列敘述何者「不正確」？ (A) 頻繁次數比之數值如果太大，表示此變數集中出現最頻繁類別，可考慮刪除此類別型變數 (B) 某類別型變數計次結果為男：2 次，女：98 次，則頻繁次數比為 98 (C) 唯一值百分比之數值太大，表示此變數幾乎完全相同，可考慮刪除此類別型變數 (D) 某類別型變數計次結果為男：2 次，女：98 次，則唯一值百分比為 2 委員釋覆結果：本題答案選項(B)與(C)均給分
D	31. 關於主成分分析（Principal Component Analysis, PCA）屬性萃取的主要用途，下列敘述何者正確？ (A) 以長條圖視覺化多變量資料 (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) 將最攸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

109 年度初級巨量資料分析師能力鑑定試題
（含疑義題釋覆結果）

科目 2：資料處理與分析概論　　疑義題釋覆結果日期:109.06.16

考試日期：109 年 5 月 30 日　　第 7 頁，共 11 頁

	變數
B	32. 關於奇異值分解（Singular Value Decomposition, SVD），下列敘述何者「不正確」？ (A) 常見應用是維度縮減 (B) 用於估計結果穩定時 (C) 用於資料其屬性個數大於觀測值個數 (D) 用於估計結果不穩定時
C	33. 根據主計總處資料，2018 年工業及服務業受僱員工（下稱受僱員工）全年總薪資如附圖，下列敘述何者「不正確」？  (https://www.managertoday.com.tw/articles/view/58998) (A) 有半數受僱員工，全年薪資不高於 49 萬元 (B) 有半數受僱員工，全年薪資不高於 62.9 萬元 (C) 此薪水分布不是左右對稱分布的鐘型曲線，而是左偏型態 (D) 全年薪資達 115 萬元，即可排名在受僱員工前 10%
C	34. 下列何者「不能」反映數據的集中趨勢的統計量？ (A) 平均數 (B) 中位數 (C) 變異數 (D) 眾數
A	35. 下列何者適合用來呈現〔車速，油耗〕資料？ (A) XY 散佈圖 (B) 直條圖 (C) 直方圖 (D) 折線圖
C	36. 關於關聯型態探勘（Association Pattern Mining），下列敘述何者「不正確」？ (A) 典型的關聯型態探勘是分析超市中顧客購買的品項集合資料

	（通常被稱為交易資料，或是購物籃資料），其個別品項間或品項群間的關聯 (B) 關聯型態挖掘最常見的模型是以品項集合出現的次數，來量化彼此間的關聯程度，以此挖掘出來的品項集合稱為頻繁品項集（Large Itemsets or Frequent Itemsets） (C) 就應用領域而言，關聯型態探勘僅應用於購物籃資料分析，因而被稱為購物籃分析（Market Basket Analysis） (D) 關聯型態探勘分析的目的是基於某些品項出現的前提下，挖掘出可預測其它品項發生之可能性的規則，這些規則就被稱為關聯規則
B	37. 自動編碼器（Autoencoder）通常「不會」用來做下列何項工作？ (A) 資料降維 (B) 無損壓縮影像 (C) 特徵擷取 (D) 去雜訊
C	38. 對於二元分類問題，依真實資料的真假值與模型預測輸出的真假值，可以組合出真陽性（True Positive, TP）、真陰性（True Negative, TN）、偽陽性（False Positive, FP）、偽陰性（False Negative, FN）四種情況，組成混淆矩陣（Confusion matrix）。若模型追求較高的精確率（precision），則應提高下列何者？ (A) TP (B) TN (C) TP/(TP+FP) (D) TP/(TP+FN)
D	39. 特徵挑選（Feature Selection）是指挑選原始資料中的合宜屬性，或可視為移除缺乏訊息內涵之變數的維度縮減策略，下列常用的降維方法中，何者屬於特徵挑選的方式？ (A) 因子分析（Factor Analysis） (B) 非負矩陣分解（Non-negative Matrix Factorization） (C) 隨機投影（Random Projections） (D) 高相關過濾（High Correlation Filter）
B	40. 特徵萃取（Feature Extraction）是指將原始資料的屬性進行結合，以產生新的代理變數（Surrogate Variables）。下列常用的降維方法中，何者屬於特徵萃取的方式？ (A) 低變異過濾（Low Variance Filter）

	(B) 多維尺度分析（Multidimensional Scaling） (C) 隨機森林（Random Forests） (D) 高相關過濾（High Correlation Filter）
C	41. 模型複雜度與預測誤差之間的變化關係，通常是越複雜的模型與訓練集合配適的越好。因此，一般而言訓練集的預測誤差，會隨著模型複雜度如何變化？ (A) 增加而增加 (B) 減少而減少 (C) 增加而減少 (D) 減少而增加
B	42. 下列何種演算法較「不適合」進行分類預測？ (A) 決策樹（Decision Tree） (B) 線性迴歸（Linear Regression） (C) 羅吉斯迴歸（Logistic Regression） (D) K-近鄰演算法（K-Nearest Neighbor）
B	43. 請問若只需輸入大學生的身高和體重來預測其腰圍，使用何種演算法較為合適？ (A) 簡單線性迴歸 (B) 多元線性迴歸 (C) 羅吉斯迴歸 (D) 關聯規則
B	44. 迴歸問題和分類問題都屬於監督式學習，關於兩者的反應變數，下列敘述何者正確？ (A) 前者是類別型反應變數，後者是數值型反應變數 (B) 前者是數值型反應變數，後者是類別型反應變數 (C) 兩者都是數值型反應變數 (D) 兩者都是類別型反應變數
A	45. 在進行機器學習時，下列何者「不是」避免過度配適（overfitting）的方法？ (A) 減少資料量 (B) 減少模型參數 (C) 使用較簡單的模型 (D) 在損失函數（loss function）加入參數權重的 L2 norm，抑制權重變大
B	46. 下列何種統計機器學習方法，容許資料中存有遺缺值？

	(A) 類神經網路（Artificial Neural Networks） (B) 分類與迴歸樹（Classification and Regression Trees） (C) K-近鄰法（K-Nearest Neighbors） (D) 羅吉斯迴歸（Logistic Regression）
B	47. 當資料集的預測變數過多時，下列哪種方法是從只有截距項的最簡單模型出發，逐步加入重要的變數？ (A) 後向式逐步迴歸 (B) 前向式逐步迴歸 (C) 中向式逐步迴歸 (D) 反覆式逐步迴歸
C	48. 關於羅吉斯迴歸（Logistic Regression）分類，下列敘述何者「不正確」？ (A) 它是建立二元類別機率值之勝率（odds ratio）對數值的線性分類模型 (B) 其反應變數假設是二項式隨機變數 (C) 它對反應變數直接建模，將之關連到預測變數的線性函數 (D) 常被人誤解成數值迴歸技術
B	49. 關於機器學習中的交叉驗證（Cross-Validation），下列敘述何者正確？ (A) 使用不同架構的模型在相同的資料上，以驗證訓練效果 (B) 是預測評估模型配適（fitting）度及尋找模型參數的方法 (C) 用來避免配適不足（underfitting） (D) 將資料分割成訓練集（training set）跟測試集（testing set），進行訓練與分析
D	50. 關於模型訓練與測試機制中的資料切分，下列敘述何者「不正確」？ (A) 實務上常用重抽樣法進行模型最佳化 (B) 決定最佳的模型複雜度或參數組合後，最後再以整個校驗集（calibration set）建立最佳複雜度或最佳參數組合下的最終模型 (C) 雙重重抽樣法包含內外兩圈的重抽樣機制，分別負責模型最佳化與績效估計的工作，如此內外圈反覆執行所需計算量應是負擔最重的訓練與測試機制 (D) 生醫或化學計量學等領域常因所搜集到樣本通常較少，因而採行 50%的訓練集（training set）用以建立模型，25%的驗證集（validation set）進行模型參數最佳化，以及 25%的測試集

	（test set）測試最終模型等三個子集的切分方式