

# 110 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：110 年 8 月 21 日

第 1 頁，共 16 頁

## 單選題 50 題 (佔 100%)

C	<p>1. 關於資料匯入與匯出，下列敘述何者「不」正確？</p> <p>(A) 針對來源資料量設計資料萃取 (Extract) 方式，例如：將大檔案切分為數個小檔案後，各別進行資料萃取作業</p> <p>(B) ETL 作業常有上下游作業關係，因此需要設定好相互關係 (Job Dependency) 與執行順序</p> <p>(C) 考慮到關聯式資料庫之 ETL 作業執行效率，可一次執行多個大資料表 (Table) 關聯 (Join)，讓資料一次寫入目的地</p> <p>(D) 當 ETL 作業發生錯誤時，規劃良好的 ETL 作業具有分階段重新執行能力 (Re-run)，不用每次都重頭開始</p>				
D	<p>2. R 語言中，使用 read.table 匯入以下特性的文字檔資料，並指派為 mydf 物件，選項中何者是符合題目要求條件之正確語法？(1)檔案名稱為「ipas.csv」、(2)資料以「,」區隔、(3)檔案編碼為「UTF-8」、(4)資料沒有標題列</p> <p>(A) <code>mydf &lt;- read.table('ipas.csv')</code></p> <p>(B) <code>mydf &lt;- read.table('ipas.csv', header = TRUE, fileEncoding = 'UTF-8')</code></p> <p>(C) <code>mydf &lt;- read.table('ipas.csv', header = TRUE, sep = ",", fileEncoding = 'UTF-8')</code></p> <p>(D) <code>mydf &lt;- read.table('ipas.csv', sep = ",", fileEncoding = 'UTF-8')</code></p>				
C	<p>3. 在製作一個 CSV (Comma-Separated Values) 檔案的過程中，如果有欄位的資料中含有逗號 (例如金額的千位號)，請問應如何處理該欄位資料最為恰當？</p> <p>(A) 透過文字編輯器，先將所有逗號(,)轉換為句號(.)</p> <p>(B) 當使用資料匯入工具時，先告知工具該列資料的欄位數即可</p> <p>(C) 使用雙引號(")包覆該欄位資料</p> <p>(D) 不用特別處理該欄位資料</p>				
C	<p>4. 透過 Python 載入 CSV 資料時，可能會使用 pandas 套件的 read_csv 函數，下列敘述何者「不」正確？(可參考附圖官方說明文件)</p> <p><code>pandas.read_csv(filepath_or_buffer, sep=';', ...)</code></p> <table border="1"> <tr> <td>Parameters:</td><td> <p><code>filepath_or_buffer</code>: str, pathlib.Path, py._path.local.LocalPath or any object with a read() method (such as a file handle or StringIO)</p> <p>The string could be a URL. Valid URL schemes include http, ftp, s3, and file. For file URLs, a host is expected. For instance, a local file could be <code>file:///localhost/path/to/table.csv</code></p> </td></tr> <tr> <td></td><td> <p><code>sep</code>: str, default ','</p> <p>Delimiter to use. If sep is None, the C engine cannot automatically detect the separator, but the Python parsing engine can, meaning the latter will be used</p> </td></tr> </table>	Parameters:	<p><code>filepath_or_buffer</code>: str, pathlib.Path, py._path.local.LocalPath or any object with a read() method (such as a file handle or StringIO)</p> <p>The string could be a URL. Valid URL schemes include http, ftp, s3, and file. For file URLs, a host is expected. For instance, a local file could be <code>file:///localhost/path/to/table.csv</code></p>		<p><code>sep</code>: str, default ','</p> <p>Delimiter to use. If sep is None, the C engine cannot automatically detect the separator, but the Python parsing engine can, meaning the latter will be used</p>
Parameters:	<p><code>filepath_or_buffer</code>: str, pathlib.Path, py._path.local.LocalPath or any object with a read() method (such as a file handle or StringIO)</p> <p>The string could be a URL. Valid URL schemes include http, ftp, s3, and file. For file URLs, a host is expected. For instance, a local file could be <code>file:///localhost/path/to/table.csv</code></p>				
	<p><code>sep</code>: str, default ','</p> <p>Delimiter to use. If sep is None, the C engine cannot automatically detect the separator, but the Python parsing engine can, meaning the latter will be used</p>				

# 110 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：110 年 8 月 21 日

第 2 頁，共 16 頁

	<div>automatically. In addition, separators longer than 1 character and different from '\s+' will be interpreted as regular expressions and will also force the use of the Python parsing engine. Note that regex delimiters are prone to ignoring quoted data. Regex example: '\r\t'</div> <div>(A) 可以載入非 CSV 格式，例如以 Tab 分隔或是以句號分隔的純文字資料</div> <div>(B) 可以直接匯入網路上的 CSV 資料，例如： pandas.read_csv('http://www.sample-videos.com/csv/Sample-Spreadsheet-10-rows.csv')</div> <div>(C) XLS（Excel Spreadsheet）資料類似於 CSV，故也可透過此函數匯入</div> <div>(D) 可以直接匯入本機資料夾的 CSV 資料</div>																																																	
B	<div>5. 關於讀取.xlsx 檔，下列敘述何者正確？</div> <div>(A) 在 R 語言中，只有一個套件{readxl}可以使用</div> <div>(B) 在 Python 語言中，可使用 pandas 套件中的 read_excel()方法</div> <div>(C) 在 R 語言中，可使用{readxl}套件中的 readxl()函數</div> <div>(D) 在 Python 語言中，可使用 pandas 套件中的 open_xlsx()方法</div>																																																	
D	<div>6. 下列何者「不」屬於 Python 物件導向特性？</div> <div>(A) 封裝（Encapsulation）</div> <div>(B) 繼承（Inheritance）</div> <div>(C) 多型（Polymorphism）</div> <div>(D) 動態配置（Dynamic Allocation）</div>																																																	
D	<div>7. 參考附圖，關於 Python 語言匯入 CSV 檔案，下列敘述何者正確？</div> <div>In [1]: import pandas as pd</div> <div>In [2]: mydata = pd.read_csv('aqx_p_434_20200626012835.csv')</div> <div>In [3]: mydata = mydata.iloc[0:6, 0:6]</div> <div>In [4]: mydata</div> <div>Out[4]:</div> <table><thead><tr><th></th><th>SiteId</th><th>SiteName</th><th>MonitorDate</th><th>AQI</th><th>S02SubIndex</th><th>COSubIndex</th></tr></thead><tbody><tr><td>0</td><td>1</td><td>基隆</td><td>2020-06-25</td><td>43</td><td>6</td><td>3.0</td></tr><tr><td>1</td><td>84</td><td>富貴角</td><td>2020-06-25</td><td>37</td><td>1</td><td>2.0</td></tr><tr><td>2</td><td>83</td><td>麥寮</td><td>2020-06-25</td><td>17</td><td>3</td><td>NaN</td></tr><tr><td>3</td><td>80</td><td>關山</td><td>2020-06-25</td><td>27</td><td>3</td><td>NaN</td></tr><tr><td>4</td><td>78</td><td>馬公</td><td>2020-06-25</td><td>15</td><td>3</td><td>2.0</td></tr><tr><td>5</td><td>77</td><td>金門</td><td>2020-06-25</td><td>23</td><td>4</td><td>1.0</td></tr></tbody></table> <div>(A) mydata 的資料筆數為 5</div> <div>(B) mydata.dropna(axis = 1)執行結果顯示 4 筆資料</div> <div>(C) sum(pd.isnull(mydata['COSubIndex']))執行結果為 3</div>		SiteId	SiteName	MonitorDate	AQI	S02SubIndex	COSubIndex	0	1	基隆	2020-06-25	43	6	3.0	1	84	富貴角	2020-06-25	37	1	2.0	2	83	麥寮	2020-06-25	17	3	NaN	3	80	關山	2020-06-25	27	3	NaN	4	78	馬公	2020-06-25	15	3	2.0	5	77	金門	2020-06-25	23	4	1.0
	SiteId	SiteName	MonitorDate	AQI	S02SubIndex	COSubIndex																																												
0	1	基隆	2020-06-25	43	6	3.0																																												
1	84	富貴角	2020-06-25	37	1	2.0																																												
2	83	麥寮	2020-06-25	17	3	NaN																																												
3	80	關山	2020-06-25	27	3	NaN																																												
4	78	馬公	2020-06-25	15	3	2.0																																												
5	77	金門	2020-06-25	23	4	1.0																																												

# 110 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：110 年 8 月 21 日

第 3 頁，共 16 頁

	(D) mydata.dropna()執行結果顯示 4 筆資料																																																																	
D	<p>8. 參考附圖，Python 語言中，選項中何者為計算各列的平均值？</p> <pre>In [1]: import pandas as pd ...: import numpy as np ...: df = pd.DataFrame([[1, 2, 3], ...:                    [4, 5, 6], ...:                    [1, np.nan, np.nan]], ...:                    columns=['x1', 'x2', 'x3'])</pre> <p>(A) df.aggregate("mean") (B) df.aggregate("mean", axis=0) (C) df.aggregate("mean", axis="index") (D) df.aggregate("mean", axis=1)</p>																																																																	
C	<p>9. 有一 pandas DataFrame 格式的變數 df，其資料內容如下：</p> <table><tr><th></th><th>姓名</th><th>科目</th><th>分數</th></tr><tr><td>1</td><td>Alice</td><td>English</td><td>75</td></tr><tr><td>2</td><td>Alice</td><td>History</td><td>80</td></tr><tr><td>3</td><td>Alice</td><td>Math</td><td>95</td></tr><tr><td>4</td><td>Ken</td><td>English</td><td>80</td></tr><tr><td>5</td><td>Ken</td><td>History</td><td>92</td></tr><tr><td>6</td><td>Ken</td><td>Math</td><td>85</td></tr><tr><td>7</td><td>Tony</td><td>English</td><td>80</td></tr><tr><td>8</td><td>Tony</td><td>History</td><td>65</td></tr><tr><td>9</td><td>Tony</td><td>Math</td><td>90</td></tr></table> <p>在執行分組彙總程式碼後（示意如下圖），得到了分組表格結果如下：</p> <pre>df.groupby("姓名").agg({     "科目": ["min", "max", "mean"] }).reset_index()</pre> <table><tr><th></th><th>(1)</th><th colspan="3">分數</th></tr><tr><th></th><th></th><th>(2)</th><th>(3)</th><th>(4)</th></tr><tr><td>0</td><td>Alice</td><td>83.333333</td><td>75</td><td>95</td></tr><tr><td>1</td><td>Ken</td><td>85.666667</td><td>80</td><td>92</td></tr><tr><td>2</td><td>Tony</td><td>78.333333</td><td>65</td><td>90</td></tr></table> <p>請問表格中(1)、(2)、(3)、(4)依序內容，下列選項何者較為符合？</p> <p>(A) 科目、mean、min、max (B) 科目、median、min、max (C) 姓名、mean、min、max (D) 姓名、median、mean、max</p>		姓名	科目	分數	1	Alice	English	75	2	Alice	History	80	3	Alice	Math	95	4	Ken	English	80	5	Ken	History	92	6	Ken	Math	85	7	Tony	English	80	8	Tony	History	65	9	Tony	Math	90		(1)	分數					(2)	(3)	(4)	0	Alice	83.333333	75	95	1	Ken	85.666667	80	92	2	Tony	78.333333	65	90
	姓名	科目	分數																																																															
1	Alice	English	75																																																															
2	Alice	History	80																																																															
3	Alice	Math	95																																																															
4	Ken	English	80																																																															
5	Ken	History	92																																																															
6	Ken	Math	85																																																															
7	Tony	English	80																																																															
8	Tony	History	65																																																															
9	Tony	Math	90																																																															
	(1)	分數																																																																
		(2)	(3)	(4)																																																														
0	Alice	83.333333	75	95																																																														
1	Ken	85.666667	80	92																																																														
2	Tony	78.333333	65	90																																																														
D	10. 有一 pandas DataFrame 格式的變數 df，其資料內容如下：																																																																	

# 110 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：110 年 8 月 21 日

第 4 頁，共 16 頁

	<table><tr><th></th><th>姓名</th><th>電話</th><th>信箱</th></tr><tr><td>0</td><td>Alfred</td><td>091AB35874</td><td>NaN</td></tr><tr><td>1</td><td>Batman</td><td>091XY35221</td><td>test01@gggmail.com</td></tr><tr><td>2</td><td>Catwoman</td><td>093XY68668</td><td>go11@gggmail.com</td></tr><tr><td>3</td><td>Chris</td><td>NaN</td><td>NaN</td></tr><tr><td>4</td><td>Ken</td><td>093XY51333</td><td>NaN</td></tr></table> <p>df_new = df.dropna(subset=['電話', '信箱'])</p> <p>print(list(df_new.姓名.unique()))</p> <p>請問執行附圖程式碼後，下列何者為輸出結果？</p> <p>(A) ['Alfred', 'Batman', 'Catwoman']</p> <p>(B) ['Alfred', 'Batman', 'Catwoman', 'Chris', 'Ken']</p> <p>(C) ['Alfred', 'Ken']</p> <p>(D) ['Batman', 'Catwoman']</p>		姓名	電話	信箱	0	Alfred	091AB35874	NaN	1	Batman	091XY35221	test01@gggmail.com	2	Catwoman	093XY68668	go11@gggmail.com	3	Chris	NaN	NaN	4	Ken	093XY51333	NaN																
	姓名	電話	信箱																																						
0	Alfred	091AB35874	NaN																																						
1	Batman	091XY35221	test01@gggmail.com																																						
2	Catwoman	093XY68668	go11@gggmail.com																																						
3	Chris	NaN	NaN																																						
4	Ken	093XY51333	NaN																																						
D	<p>11. 參考附圖，R 語言中，下列邏輯值索引敘述何者正確？</p> <pre>&gt; mydata &lt;- c(-10^-3, -10^-2, 0, 10^2, 10^3) &gt; mydata [1] -1e-03 -1e-02 0e+00 1e+02 1e+03 &gt;</pre> <p>(A) mydata[-1]執行結果為 1e+03</p> <p>(B) sum(mydata)執行結果為 0</p> <p>(C) mydata[0]執行結果為-1e-03</p> <p>(D) mydata[3]執行結果為 0</p>																																								
B	<p>12. 有一 pandas DataFrame 格式的變數 df，其資料內容如下：</p> <table><tr><th></th><th>姓名</th><th>已就職滿月數</th><th>上個月請假日數</th></tr><tr><td>0</td><td>Ted</td><td>12</td><td>2</td></tr><tr><td>1</td><td>Jen</td><td>3</td><td>0</td></tr><tr><td>2</td><td>Peter</td><td>13</td><td>3</td></tr><tr><td>3</td><td>Ninn</td><td>3</td><td>0</td></tr><tr><td>4</td><td>Celine</td><td>10</td><td>1</td></tr><tr><td>5</td><td>Judy</td><td>1</td><td>0</td></tr><tr><td>6</td><td>Wendy</td><td>15</td><td>2</td></tr><tr><td>7</td><td>Andrew</td><td>9</td><td>0</td></tr><tr><td>8</td><td>Shawn</td><td>12</td><td>0</td></tr></table> <pre>def check_isValid(row):     ''' 判斷是否可領全勤獎金 '''     rName = row["姓名"]     rMonth = row["已就職滿月數"]     rLeaveDays = row["上個月請假日數"]     if rMonth &gt;= 3 and rLeaveDays &lt;= 1:</pre>		姓名	已就職滿月數	上個月請假日數	0	Ted	12	2	1	Jen	3	0	2	Peter	13	3	3	Ninn	3	0	4	Celine	10	1	5	Judy	1	0	6	Wendy	15	2	7	Andrew	9	0	8	Shawn	12	0
	姓名	已就職滿月數	上個月請假日數																																						
0	Ted	12	2																																						
1	Jen	3	0																																						
2	Peter	13	3																																						
3	Ninn	3	0																																						
4	Celine	10	1																																						
5	Judy	1	0																																						
6	Wendy	15	2																																						
7	Andrew	9	0																																						
8	Shawn	12	0																																						

# 110 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：110 年 8 月 21 日

第 5 頁，共 16 頁

	<pre>return 1 return 0  df['是否可領上個月全勤獎金'] = df.apply(check_isValid, axis=1) bonusNameList = list(df.query("是否可領上個月全勤獎金 == 1").姓名.unique()) print(bonusNameList)</pre> <p>請問執行附圖程式碼後，下列選項內姓名何者「不」在執行結果當中？</p> <p>(A) Andrew (B) Judy (C) Celine (D) Jen</p>																																				
C	<p>13. 有一 pandas DataFrame 格式的變數 df，其資料內容如下：</p> <table><tr><th></th><th>姓名</th><th>班級</th><th>國文</th><th>英文</th><th>數學</th></tr><tr><td>0</td><td>John</td><td>A</td><td>90</td><td>80</td><td>60</td></tr><tr><td>1</td><td>Ken</td><td>A</td><td>70</td><td>80</td><td>90</td></tr><tr><td>2</td><td>Norman</td><td>B</td><td>100</td><td>70</td><td>90</td></tr><tr><td>3</td><td>Andy</td><td>B</td><td>80</td><td>80</td><td>80</td></tr><tr><td>4</td><td>Mona</td><td>A</td><td>90</td><td>90</td><td>60</td></tr></table> <p>關於使用 pandas 語法對變數 df 進行運算之語法與其結果，請問下列敘述何者「不」正確？</p> <p>(A) df.groupby("班級").agg({"姓名": "nunique"})，可獲得各班級人數之分組統計</p> <p>(B) df["加權分數"] = df.國文*1 + df.英文*2 + df.數學*2，可新增"加權分數"欄位至 df</p> <p>(C) df.數學.median()，可得到 df 數學欄位之平均值為 76</p> <p>(D) df[df.姓名.str.startswith("K")], 可篩選出 df 姓名開頭為 K 的資料表</p>		姓名	班級	國文	英文	數學	0	John	A	90	80	60	1	Ken	A	70	80	90	2	Norman	B	100	70	90	3	Andy	B	80	80	80	4	Mona	A	90	90	60
	姓名	班級	國文	英文	數學																																
0	John	A	90	80	60																																
1	Ken	A	70	80	90																																
2	Norman	B	100	70	90																																
3	Andy	B	80	80	80																																
4	Mona	A	90	90	60																																
C	<p>14. 有一 pandas DataFrame 格式的變數 df1，其資料內容如下：</p>																																				

# 110 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：110 年 8 月 21 日

第 6 頁，共 16 頁

A	
0	1
1	2
2	3

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

A	
0	4
1	5

若要透過 pandas 語法將 df1, df2 進行合併，產出另一 pandas DataFrame 格式的變數 df3 (df3 之數據如附圖)。請問下列哪一個選項中的語法可以產出此結果？

A	
0	1
1	2
2	3
3	4
4	5

- (A) `df3 = df1.merge(df2, on="A")`
- (B) `df3 = df1 + df2`
- (C) `df3 = pd.concat([df1, df2]).reset_index(drop=True)`
- (D) `df3 = df1.concat(df2).reset_index(drop=True)`

D 15. 附圖為某商店其中一位客戶的消費紀錄，請問下列何種資料處理方式「最不」適當？

ID	時間	消費金額	商品
1	21-04-29	507	A 商品
2	21-05-18	919	E 商品
3	2021-06-22	215	NA (遺缺值)
4	2021-07-02	339	R 商品

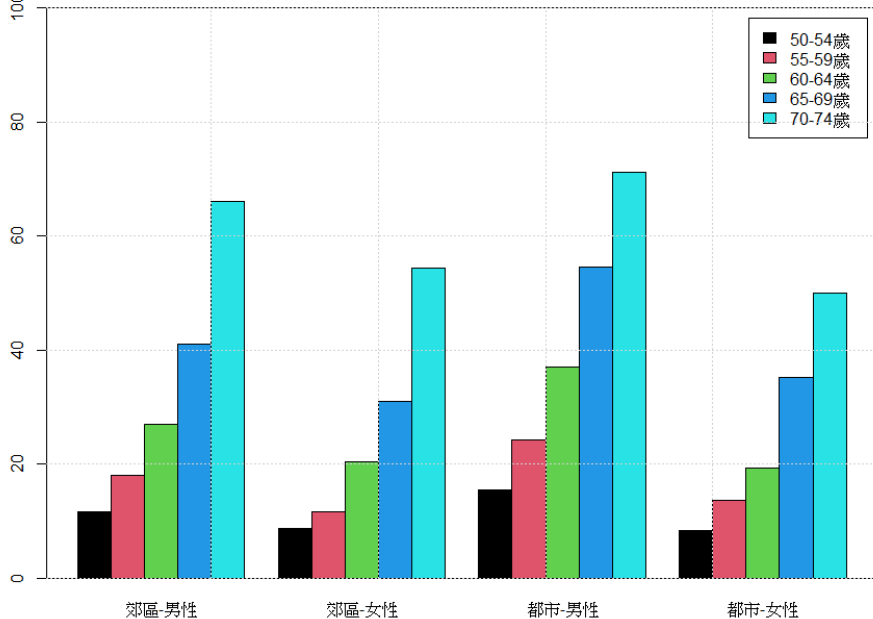
- (A) 修正錯誤的時間，建立消費金額的時間序列
- (B) 不修正時間，依照資料順序建立消費金額的時間序列
- (C) 依照消費金額尋找可能之商品名稱填補 NA 值
- (D) 移除含有 NA 值的該筆觀測值資料

# 110 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：110 年 8 月 21 日

第 7 頁，共 16 頁

A	<p>16. 下列何者「不」屬於 Python 或 R 常用到的繪圖套件？</p> <p>(A) D3.js (Data-Driven Documents)</p> <p>(B) Matplotlib</p> <p>(C) ggplot2</p> <p>(D) Seaborn</p>																														
C	<p>17. 參考附圖，R 語言中下列敘述何者正確？</p> <div><p>每千人死亡率-1940年</p><table border="1"><thead><tr><th>Category</th><th>50-54歲</th><th>55-59歲</th><th>60-64歲</th><th>65-69歲</th><th>70-74歲</th></tr></thead><tbody><tr><td>郊區-男性</td><td>12</td><td>18</td><td>28</td><td>42</td><td>68</td></tr><tr><td>郊區-女性</td><td>10</td><td>12</td><td>20</td><td>32</td><td>55</td></tr><tr><td>都市-男性</td><td>15</td><td>25</td><td>38</td><td>55</td><td>72</td></tr><tr><td>都市-女性</td><td>10</td><td>15</td><td>20</td><td>35</td><td>50</td></tr></tbody></table></div> <p>(A) 資料共有 25 筆</p> <p>(B) 四大類別中，合計死亡率最高者為「都市-女性」</p> <p>(C) 四大類別中，60-64 歲人口死亡率第二高者為「郊區-男性」</p> <p>(D) 使用 barplot 繪圖時，須設定 beside = FALSE 才可繪製附圖結果</p>	Category	50-54歲	55-59歲	60-64歲	65-69歲	70-74歲	郊區-男性	12	18	28	42	68	郊區-女性	10	12	20	32	55	都市-男性	15	25	38	55	72	都市-女性	10	15	20	35	50
Category	50-54歲	55-59歲	60-64歲	65-69歲	70-74歲																										
郊區-男性	12	18	28	42	68																										
郊區-女性	10	12	20	32	55																										
都市-男性	15	25	38	55	72																										
都市-女性	10	15	20	35	50																										
D	<p>18. 參考附圖，關於 mtcars 資料集散佈圖矩陣，下列敘述何者為正確？</p>																														

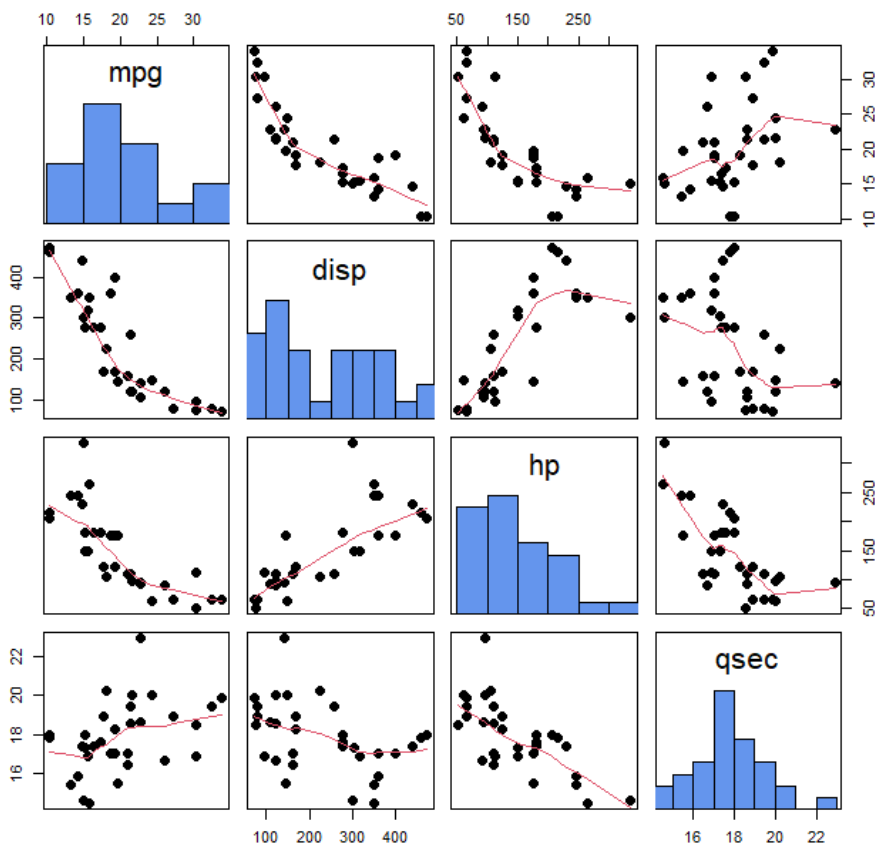


# 110 年度中級巨量資料分析師能力鑑定試題

科目 1：資料分析與資料科學

考試日期：110 年 8 月 21 日

第 8 頁，共 16 頁

	<p><b>mtcars資料集散佈圖矩陣</b></p>  <p>(A) mpg 與 disp 資料呈現約為正相關          (B) disp 與 hp 資料呈現約為負相關          (C) hp 達到 300 以上佔較大的比例          (D) qsec 與 disp 呈現約為負相關</p>
C	<p>19. 關於各類統計圖的使用時機，下列敘述何者「不」正確？</p> <p>(A) 雷達圖（Radar Char）可以用來比較多個指標          (B) 散佈圖（Scatter Plot）可以用來觀察兩連續變數 X 與 Y 之間的關係          (C) 散佈圖（Scatter Plot）可以觀察兩個變數間的因果關係          (D) 泡泡圖（Bubble），用泡泡大小展現第三變量</p>
C	<p>20. 附圖為某餐廳之“餐廳小費統計數據”所繪製而成的箱型圖（Box Plot）。關於該數據與圖表的敘述，下列何者較「不」正確？（單位：元）</p>

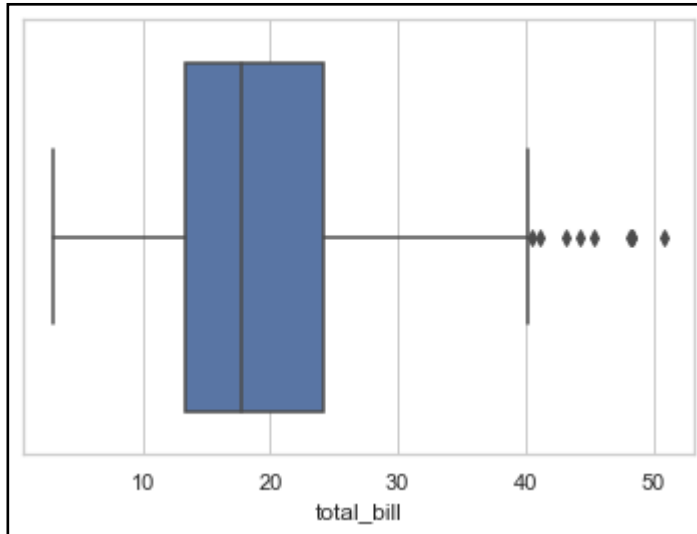


# 110 年度中級巨量資料分析師能力鑑定試題

科目 1：資料分析與資料科學

考試日期：110 年 8 月 21 日

第 9 頁，共 16 頁



- (A) 該數據之最小值大於 0 元
- (B) 該數據之中位數介於 10-20 元之間
- (C) 該數據之箱型圖繪製結果中，沒看到有離群值的情況
- (D) 該數據之第三四分位數（第 75 個百分位數）大約是 24 元

A 21. 請問資訊增益（Information Gain）衡量是應用在樹狀模型（如：決策樹）建構過程中的哪一個階段？

- (A) 分割資料集的預測變數（Attribute Selection Measures）與其分割值
- (B) 樹的深度或複雜度
- (C) 葉節點（Leaf Node）的預測方程或方式
- (D) 訓練樣本數

D 22. 關於模型績效評量，下列敘述何者「不」正確？

- (A) 沒有衡量就無法管控，任何預測模型只有運用適當的指標，評核其模型績效後方能合理的運用之
- (B) 以迴歸模型來說，許多績效評量的計算是基於殘差（Residual），或稱預測誤差（Prediction Error），也可簡稱為誤差（Error）
- (C) 均方預測誤差（Mean Squared Error, MSE）或簡稱為均方誤差，它是殘差平方值的算術平均，因其單位是原始反應變數單位的平方，容易造成數據解讀上的困擾
- (D) 衡量該模型績效的方式通常很多，實務上建議以單一評估指標瞭解特定模型的優缺點

B 23. 混淆矩陣（Confusion Matrix）是等長的觀測類別值向量與預測類別值向量，交叉統計後的二維表格結果，請問下列敘述何者「不」正確？

# 110 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：110 年 8 月 21 日

第 10 頁，共 16 頁

	<table><tr><td></td><td colspan="3">真實類別</td></tr><tr><td rowspan="3">預測類別</td><td></td><td>Cp(relevant)</td><td>Cn(not relevant)</td></tr><tr><td>Cp(retrieved)</td><td>真陽數 True Positive(TP)</td><td>假陽數 False Positive(FP)</td></tr><tr><td>Cn(not retrieved)</td><td>假陰數 False Negative (FN)</td><td>真陰數 True Negative (TN)</td></tr></table>		真實類別			預測類別		Cp(relevant)	Cn(not relevant)	Cp(retrieved)	真陽數 True Positive(TP)	假陽數 False Positive(FP)	Cn(not retrieved)	假陰數 False Negative (FN)	真陰數 True Negative (TN)
	真實類別														
預測類別		Cp(relevant)	Cn(not relevant)												
	Cp(retrieved)	真陽數 True Positive(TP)	假陽數 False Positive(FP)												
	Cn(not retrieved)	假陰數 False Negative (FN)	真陰數 True Negative (TN)												
	<p>(A) 形容詞真與假意指預測的結果是否與其真實的類別相同</p> <p>(B) 混淆矩陣中真陽數一定大於真陰數</p> <p>(C) R 語言 caret 套件與 Python 語言 pandas_ml 套件中有分類模型的各種績效評估指標</p> <p>(D) 陽性事件通常是我們所關心的事件，例如:授信客戶違約、垃圾郵件與簡訊、患有某種疾病等，當這些事件發生時，人們通常會採取因應措施</p>														
B	<p>24. 交叉驗證 (Cross-Validation) 主要用於模型訓練或建模應用中，目的是為了得到可靠穩定的模型。請問下列敘述何者正確？</p> <p>(A) k 摺交叉驗證 (k-fold Cross Validation)，若 k=10，代表將數據集分成 10 份，將其中 5 份做訓練、5 份做驗證</p> <p>(B) 交叉驗證經常用於分類預測、偏最小平方 (Partial Least Squares, PLS) 迴歸建模等</p> <p>(C) 採用 k 摺交叉驗證 (k-fold Cross Validation) 通常會重複 k 次以上，以 k-1 次的結果均值作為對算法精度的估計</p> <p>(D) 保留法 (Holdout) 驗證不算交叉驗證的類型</p>														
C	<p>25. 拔靴集成法(Bootstrap AGGREGatING, BAGGING)是一種常見的重抽樣 (Resampling) 方法，下列敘述何者「不」正確？</p> <p>(A) 拔靴集成法是將已有的觀察值當作是母體，重複進行抽樣</p> <p>(B) 當有充足的樣本數、且樣本具有與母體類似的特性時，拔靴集成法可用來近似分配的形狀</p> <p>(C) 拔靴集成法常用於大量數據資料重抽樣</p> <p>(D) 拔靴集成法可以作為交叉驗證的一個替代方法，在原本的樣本中進行替換的隨機採樣，從而得到新的樣本</p>														
A	<p>26. 真實的反應變數值與預測的反應變數值之間的差，稱為殘差或 (預測) 誤差，下列何者是應用殘差平方值的總和來評估迴歸模型的績效？</p> <p>(A) 誤差平方和 (Sum of the Squared Errors, SSE)</p>														

# 110 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：110 年 8 月 21 日

第 11 頁，共 16 頁

	<p>(B) 均方預測誤差 (Mean Squared Error, MSE)</p> <p>(C) 均方根預測誤差 (Root Mean Squared Error, RMSE)</p> <p>(D) 誤差絕對值和 (Sum of Absolute Error, SAE)</p>
C	<p>27. 建構決策樹過程中常以資訊增益 (Information Gain) 為分類的指標，請問資訊增益是以下列何者為基礎？</p> <p>(A) 變異數</p> <p>(B) 四分位距</p> <p>(C) 熵 (Entropy) 係數</p> <p>(D) 中位數絕對離差</p>
B	<p>28. 當資料科學家建模時，下列何者為過度配適 (Over-fitting) 的狀況？</p> <p>(A) 測試誤差高，訓練誤差高</p> <p>(B) 測試誤差高，訓練誤差低</p> <p>(C) 測試誤差低，訓練誤差低</p> <p>(D) 測試誤差低，訓練誤差高</p>
D	<p>29. 下列選項何者為梯度下降法的正確步驟順序？(1)重複迭代，直到得到權重最佳值、(2)把輸入傳入類神經網路，得到輸出、(3)對每一個神經元計算誤差與調整相對應的權重以減少誤差、(4)用隨機值初始化權重與偏差、(5)計算預測值與真實值之間的誤差</p> <p>(A) 45132</p> <p>(B) 24513</p> <p>(C) 45312</p> <p>(D) 42531</p>
D	<p>30. 機器學習模型中，關於模型的偏差 (Bias) 與變異 (Variance)，下列敘述何者正確？</p> <p>(A) 高偏差代表模型過於複雜</p> <p>(B) 高變異代表模型過於簡單</p> <p>(C) 模型訓練的目標為低偏差與高變異</p> <p>(D) 偏差與變異之間存在平衡 (Trade-off) 關係</p>
B	<p>31. 若執行新專案，輸入的向量超過 100 維，下列何種方法「不」適合用來做為降維並取得特徵向量的方法？</p> <p>(A) PCA (Principal Component Analysis)</p> <p>(B) LSTM (Long Short-Term Memory)</p> <p>(C) ICA (Independent Component Analysis)</p> <p>(D) Autoencoder</p>
D	<p>32. 附圖是某次瑕疵檢測的混淆矩陣，此次總共檢測 100 片電路板，實際有瑕疵的電路板有 9 片，下列敘述何者「不」正確？</p>

# 110 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：110 年 8 月 21 日

第 12 頁，共 16 頁

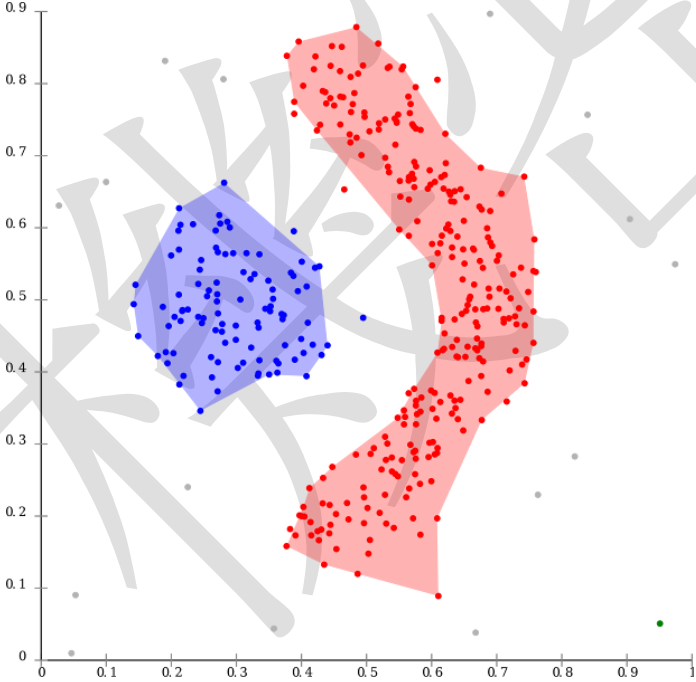
	<table><tr><th>預測 \ 實際</th><th>True</th><th>False</th></tr><tr><th>True</th><td>1</td><td>1</td></tr><tr><th>False</th><td>8</td><td>90</td></tr></table> <p>(A) 正確率 (Accuracy) 為 0.91</p> <p>(B) F1-measure 為 0.18</p> <p>(C) Recall Rate 為 0.11</p> <p>(D) Precision Rate 為 0.11</p>	預測 \ 實際	True	False	True	1	1	False	8	90
預測 \ 實際	True	False								
True	1	1								
False	8	90								
C	<p>33. 基於集群的離群值之偵測方法 (Clustering-based Approaches)，下列何者「不」是用來判斷離群值的依據？</p> <p>(A) 物件不屬於任何集群</p> <p>(B) 物件與最接近的集群之間是否存在較大距離</p> <p>(C) 物件是否位於兩個集群之間</p> <p>(D) 物件是小型或稀疏集群的一部分</p>									
B	<p>34. 關於資料解析，下列敘述何者「不」正確？</p> <p>(A) 特徵工程 (Feature Engineering) 是一種資料解析的過程，使用專業背景知識和技巧處理數據，使得特徵能在機器學習算法上發揮更好的作用</p> <p>(B) 資料解析過程中，若遇到有異常的數據資料，考慮時效性應該直接予以剔除</p> <p>(C) 資料變數篩選，可以採取專家討論或使用分析方法的方式來進行</p> <p>(D) 資料整理、解析、變數篩選等步驟，往往佔據建模過程中大量的時間</p>									
D	<p>35. 建模的過程中，經常會出現不平衡資料 (Imbalanced Data) 的問題，下列敘述何者「不」正確？</p> <p>(A) 採用數據合成，例如：SMOTE (Synthetic Minority Oversampling Technique)</p> <p>(B) 使用採樣，例如：上採樣 (Oversampling)、下採樣 (Undersampling)</p> <p>(C) 採用加權方式處理</p> <p>(D) 採用梯度下降法 (Gradient Descent)</p>									
A	<p>36. 關於特徵工程 (Feature Engineering)，下列敘述何者正確？</p> <p>(A) 特徵交叉是一種很獨特的方式，它將兩個或更多的類別屬性組合成一個，當組合的特徵要比單個特徵更好時，這是一項非常有用的技術</p> <p>(B) 時間戳屬性通常只需要分離成一兩個維度，比如：年、月，其他</p>									

# 110 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：110 年 8 月 21 日

第 13 頁，共 16 頁

	<p>太細如：日、小時、分鐘、秒鐘等就不需要了</p> <p>(C) 遇到類別型屬性的資料，不可以採用單熱編碼（One-hot Encoding）方式來進行分解</p> <p>(D) 逐步迴歸經常用於特徵縮放</p>
A	<p>37. 貝氏定理主要是哪三種機率構成？</p> <p>(A) 事前機率、事後機率、條件機率</p> <p>(B) 事前機率、聯合機率、條件機率</p> <p>(C) 事前機率、獨立事件、條件機率</p> <p>(D) 聯合機率、獨立事件、條件機率</p>
A	<p>38. 典型的 k 平均數（k-means）屬於下列何種集群（Clustering）方式？</p> <p>(A) 分割式集群（Partitional Clustering）</p> <p>(B) 階層式集群（Hierarchical Clustering）</p> <p>(C) 密度集群（Density-based Clustering）</p> <p>(D) 基於圖的集群（Graph-based Clustering）</p>
C	<p>39. 請問附圖最可能是使用何種分群方法的結果？</p>  <p>(圖取自維基百科)</p> <p>(A) 分割式集群（Partitional Clustering）</p> <p>(B) 階層式集群（Hierarchical Clustering）</p> <p>(C) 密度集群（Density-based Clustering）</p> <p>(D) 基於圖的集群（Graph-based Clustering）</p>
B	<p>40. 機器學習建模的過程中，面對解決模型變異過高的問題，下列何者是添加建模過程的隨機噪訊，以有效降低模型變異的方法？</p> <p>(A) 生成式對抗網路（Generative Adversarial Networks, GAN）</p>



# 110 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：110 年 8 月 21 日

第 14 頁，共 16 頁

	<p>(B) 拔靴集成法 (Bootstrap AGGREGatING, BAGGING)</p> <p>(C) 支援向量機 (Support Vector Machines, SVM)</p> <p>(D) 策略梯度方法 (Policy Gradient Method)</p>
C	<p>41. 在集成學習 (Ensemble Learning) 中，拔靴集成法 (Bootstrap AGGREGatING, BAGGING) 和提升法 (Boosting) 是兩種常見的技術，關於兩者的比較，下列敘述何者正確？</p> <p>(A) 提升法解決了拔靴集成法的過度配適 (Over-fitting) 問題，因此有較好的分類準確率</p> <p>(B) 拔靴集成法會根據每個樣本的重要性不同，調整不同權重，而提升法中的每個樣本的權重皆相同</p> <p>(C) 提升法需要依序訓練各個分類器，拔靴集成法則可以平行訓練各個分類器</p> <p>(D) 拔靴集成法會產生袋外 (Out-of-bag) 資料，但提升法不會</p>
C	<p>42. 關於隨機森林 (Random Forest)，下列敘述何者正確？</p> <p>(A) 處理的問題涉及序列相關的決策 (Sequential Decisions)</p> <p>(B) 產生的模型集合俗稱裝袋樹 (Bagged Trees)</p> <p>(C) 在模型中融入屬性隨機挑選的機制</p> <p>(D) 經常使用在 Kaggle 競賽中的統計機器學習算法之一，以建立多個互補的弱模型 (Weak Learner) 提升效能</p>
B	<p>43. 在隨機森林演算法中，如果資料數目總共為 <math>N</math> 個，該如何進行拔靴集成法 (Bootstrap AGGREGatING, BAGGING) 處理？</p> <p>(A) 從 <math>N</math> 個資料分布中，挑選位於平均正負一個標準差的樣本，並將樣本放回</p> <p>(B) 從 <math>N</math> 個資料中取 <math>n</math> 個資料並將樣本放回</p> <p>(C) 從 <math>N</math> 個資料分布中，挑選位於平均正負一個標準差的樣本，並不將樣本放回</p> <p>(D) 從 <math>N</math> 個資料中取 <math>n</math> 個資料並且不將樣本放回</p>
C	<p>44. 關於處理不平衡的數據資料集，下列何者「不」是常見採用的解決方法？</p> <p>(A) <math>k</math> 折交叉驗證法 (<math>k</math>-fold Cross-validation)</p> <p>(B) 數據複製 (Repetition)</p> <p>(C) C5.0 法</p> <p>(D) 拔靴法 (Bootstrapping)</p>
B	<p>45. 關於效能提升法 (Boosting)，下列敘述何者「不」正確？</p> <p>(A) 是集成學習 (Ensemble Learning) 的一種方法</p> <p>(B) 可以用來減少變異數 (Variance)</p> <p>(C) 自適應 Boosting (AdaBoost) 是一種改良效能提升的方法</p>

# 110 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：110 年 8 月 21 日

第 15 頁，共 16 頁

	(D) 自適應 Boosting (AdaBoost) 方法中的各個學習器存在著強依賴關係
B	<p>46. 關於隨機森林 (Random Forest) 的建立過程中，下列敘述何者「不」正確？</p> <p>(A) 因為隨機採樣的關係，就算不剪枝，也較不會出現過度配適 (Over-fitting) 的現象</p> <p>(B) 隨機森林是對決策樹 (Decision Tree) 的一種改進，森林中的每棵樹具有不同的分佈</p> <p>(C) 當隨機森林中的決策樹個數很多時，進行資料訓練時需要的空間和時間會比較大</p> <p>(D) 隨機森林能處理很高維度的資料，並且不用做特徵篩選</p>
D	<p>47. 下列何者「不」是 k 平均數 (k-means) 集群法的特點？</p> <p>(A) 算法涉及隨機抽樣，每次運行的結果不盡相同</p> <p>(B) 原理簡單，容易以非統計的詞彙解釋說明之</p> <p>(C) 形成的群多為類圓球狀且大小相近</p> <p>(D) 不易受到離群值的影響</p>
C	<p>48. 關於關聯分析之 FP-growth (Frequent Pattern-growth) 演算法，下列敘述何者正確？</p> <p>(A) 這是一種需要生成候選項目集的頻繁項目集探勘方法</p> <p>(B) 採用類似 Apriori 方法的生成和測試 (Generate-and-test) 策略</p> <p>(C) 構造了一個高度緊湊的資料結構 (FP-tree) 來壓縮原始交易資料庫</p> <p>(D) 著重於多次掃描資料庫以避免昂貴的候選生成</p>
A	<p>49. 關於非監督式學習 (Unsupervised Learning)，下列敘述何者正確？</p> <p>(A) 在訓練時僅須對機器提供輸入範例，非監督式學習的方法會自動從這些範例中找出潛在的規則</p> <p>(B) KNN (K Nearest Neighbor) 演算法屬於非監督式學習方法</p> <p>(C) 針對網站上線後進行 A/B Test 是屬於非監督式學習的一種實務應用</p> <p>(D) 因為對大量資料進行標籤相當費時，所以非監督式學習只需要對少部分資料進行標籤即可</p>
D	<p>50. 關於支援向量機 (Support Vector Machines, SVM)，下列敘述何者「不」正確？</p> <p>(A) 是分類、異常偵測與迴歸的工具</p> <p>(B) 藉由最大化超平面與資料之間的邊界幅度，決定分割步同樣本的最佳決策邊界</p> <p>(C) 以原始空間內積，來表達屬性空間中向量的內積，而計算屬性空</p>



# 110 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：110 年 8 月 21 日

第 16 頁，共 16 頁

	間中向量內積的函數稱為核函數 (Kernel Function) (D) 易受雜訊影響，容易過度配適 (Over-fitting)
--	--

機密