BCS Level 4 Module in Data Analysis

Sample Paper A

Record your surname / last / family name and initials on the answer sheet.

Sample paper only consisting of 20 questions in total across:

- 10 knowledge questions that include a range of question types such as multiple choice, multiple response, fill in the blanks and ordering question types – 1 mark awarded for each question.
- 2 skills scenarios each with 5 questions designed to test knowledge, skills and behaviours that include a range of question types such as multiple choice, multiple response, fill in the blanks and ordering question types – 1 mark awarded for each question.

A number of possible answers are given for each multiple choice or multiple response question, indicated by either A B C or D (up to E in the skills scenarios). A number of other questions will require you to re-order a list or fill in the blanks. Your answers should be clearly indicated on the answer sheet.

Pass mark is 13/20
Time allowed: 45 minutes

Copying of this paper is expressly forbidden without the direct approval of BCS, The Chartered Institute for IT.

This professional certification is not regulated by the following United Kingdom Regulators - Ofqual, Qualifications in Wales, CCEA or SQA.
1 Which of the following would be described as proprietary data?
A Internally generated documents relating to an organisation’s research and development activities.
B Government census results.
C Data available to anyone with a possibility of redistribution without copyright restriction.
D Documents an organisation wishes to keep confidential and not available to its competitors.

2 Order the following into the correct sequence of the Data Lifecycle.
A Use
B Store
C Delete
D Create
E Archive

3 Which of the following is NOT a TRUE statement relating to a relational database?
A They can be accessed with SQL.
B They often use foreign / primary keys.
C Their structure can be normalised.
D They cannot contain more than one single table.

4 Which of the following is NOT a source of unstructured data?
A A video file held within a folder.
B A list of customers with their address.
C A set of posts made via social media.
D A handwritten story.

5 What type of analytics summarises what has happened in the past?
A Prescriptive.
B Decision.
C Descriptive.
D Predictive.
6 Which is **NOT** an objective of the latest Data Protection Act?

A. Give Government more control over what data is being held.
B. To prevent organisations holding inaccurate information.
C. To give people more control over how their data is handled.
D. To be able to delete data when it is no longer relevant.

7 Some business systems will deliberately reduce the quality of the data by omitting or obscuring personal information. What would be a good reason for doing this?

A. It preserves customer confidentiality.
B. It makes the conceptual model simpler.
C. It increases the data accuracy.
D. It avoids costly data backups.

8 Which of the following is a widely recognized term for managing an organisation’s critical data, providing a single point of reference?

A. Important Data Management.
B. Master Data Management.
C. Vital Data Management.
D. Essential Data Management.

9 Which of the following best describes the purpose of hypothesis testing?

A. To interpret whether our null hypothesis is significantly likely to happen or whether this hypothesis should be rejected.
B. To prove that, with two competing hypotheses, the alternative hypothesis population parameter is true.
C. To prove that, with two competing hypotheses, that one is more likely than another.
D. To find the correlation between two related sets of data.

10 Which of the following chart types is **MOST** suitable for categorical data?

A. Scatter plot.
B. Radar chart.
C. Line graph.
D. Pie chart.
Scenario 1: Database design and SQL

You have been put in charge of running the monthly scorecard for the IT department. You will be using the 'IT Assets' database to retrieve the relevant data. The main tables to retrieve information from are shown in the ERD below.

More information is required on employees to help escalate issues with IT asset conditions. It has been decided to add another table storing information on an employee's manager.

What new field could be a suitable primary key for the Manager table?

A. first_name
B. last_name
C. department
D. date_started
E. manager_id
12 Which field in the Employees table would be a suitable foreign key in the new 'Managers' table?

A first_name  
B last_name  
C department  
D employee_id  
E other_details

13 A key report measure on the scorecard is the number of assets returned in a month and the condition when returned. Which of the following aggregations should be used to query the 'Employee_Assets' table for this measure?

A Group By  
B Count  
C Sort Ascending  
D Left  
E Union

14 If an asset_id is removed from the IT_assets table, all related information for that asset should also be removed from the other tables. Which database functionality should be relied upon in this instance?

A Delete query  
B Update query  
C Cascading Delete  
D Index  
E Date fields

15 Fill in the blanks to complete the SQL query shown below, using the given options listed.

The report requires a view of how many employees are not entering their name when taking an asset. This report should show how many employees per department are leaving their name blank.

```
SELECT _____(employee_id), department from ______ where first_name = ""
group by ______ order by ________
```

Blank options: department, EMPLOYEES, COUNT, MAX, 2, 3
Scenario 2 - Data Preparation and Integration

You are working as a data analyst in the IT department and have been asked to look into processing and analysing the IT operational data. The IT department are keen to operationalise the reporting processes and are therefore using programming languages to automate the importing, cleansing and manipulation of the data.

16 Order the following Python commands into a logical flow for importing the data contained in the file called "IToutages.csv".

After importing the entire file you should print the first ten characters of data to screen.

```python
f.close()
print(FirstTenChars)
data=f.read()
f = open("IToutages.csv")
FirstTenChars = data[0:9]
```

You can use this space below to provide your answer unless using a separate answer sheet:

17 Which is the missing line of code in the following python programme to find the mean of 2,3,4,5,6?

```python
Numbers = [2,3,4,5,6]
Total=sum(Numbers)
print(Mean)
```

A Mean=2*6/4
B Range=6-2
C Mean=Total/5
D Mean=sqrt(Total)
E Mean=Total**2

18 Which of the following R commands will correctly show different averages and quartiles of a dataset?

A str(dataset)
B quantile(dataset)
C mean(dataset)
D summary(dataset)
E median(dataset)
19  The IT department are also considering using the R language. What R command could you use to see a snapshot of the first six rows in a dataset?

A  head(dataset)
B  head(dataset, 5)
C  tail(dataset,6)
D  dataset[6]
E  dataset[5]

20  You should be able to visualise time series data very quickly in R. Order the following lines of code into a logical flow to read in the 'ITemployees.txt' file and plot the data as an annual time series. Having realised that the data is not annual you should then change the timeseries to be monthly and the plot a second graph that starts in 1999.

A  birthstimeseries <- ts(BirthsByYear)
B  monthplot(birthstimeseries)
C  BirthsByYear <- read.csv("births.txt")
D  plot.ts(birthstimeseries)
E  birthstimeseries <- ts(BirthsByYear, frequency = 12, start = 1999)

End of Paper
<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>Explanation / Rationale</th>
<th>Syllabus Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A and D</td>
<td>Proprietary data is internally generated data or documents that contain technical or other types of information controlled by a firm to safeguard its competitive edge. It may be protected under copyright or trade secret laws.</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>See explanation</td>
<td>Create. Store. Use. Archive. Delete.</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
<td>Relational Databases are not restricted to containing only a single table. The relationships of relational databases are generally logical connections between multiple tables.</td>
<td>2.2</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>Structured data is that which can be easily sorted in a tabular format and is often organised in a relational database or spreadsheet using fixed fields and columns. Due to the highly organised nature of structured data it is easily searchable. Examples of structured data include dates, names, location, sales figures, or even population numbers.</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>Descriptive analytics is a preliminary stage of data processing that creates a summary of historical data.</td>
<td>3.2</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>The government is not the holder of all information and there is nothing in the act to increase data held by government.</td>
<td>3.4</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>Obscuring or removing personal data will protect the individual by preventing accurate, genuine data from being available.</td>
<td>3.9</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>In business, Master Data Management (MDM) comprises the processes, governance, policies, standards and tools that consistently define and manage the critical data of an organisation to provide a single point of reference.</td>
<td>4.3</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that there is nothing ‘happening’ (null hypothesis) or whether this can be rejected to accept the alternative.</td>
<td>6.1</td>
</tr>
<tr>
<td>10</td>
<td>D</td>
<td>A pie chart is an accurate graphical representation of the categorical distribution of data.</td>
<td>6.4</td>
</tr>
<tr>
<td>Question</td>
<td>Answer</td>
<td>Explanation / Rationale</td>
<td>Syllabus Section</td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>-------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>11</td>
<td>E</td>
<td>The answer field implies an 'id' based on manager which should lead through thought process to unique identifier of manager.</td>
<td>4.4, 4.6</td>
</tr>
<tr>
<td>12</td>
<td>D</td>
<td>Employee_id is the primary key in the linked table so the only option to be described as a foreign key. Whilst alternatives could be argued, this is a case of the only sensible option from the ones given.</td>
<td>4.4, 4.6</td>
</tr>
<tr>
<td>13</td>
<td>A and B</td>
<td>To return results from a query that gives the number of assets will mandate use of a 'Count' aggregation. As the 'Count' aggregation won't be required on the rest of the data i.e. condition, this will need to be aggregated using a 'Group By'</td>
<td>4.7</td>
</tr>
<tr>
<td>14</td>
<td>C</td>
<td>This question tests deeper database functionality rather than SQL. The feature of 'cascading delete' is standard functionality across different systems and would be used to ensure related records are removed from related tables.</td>
<td>4.7</td>
</tr>
<tr>
<td>15</td>
<td>See explanation</td>
<td>The answer would require choice of options (four from six) that return the number of employees (count) using the criteria of blank first name. The group by will require the department name to match the question criteria and the order by checks understanding of labelling columns rather than using field names (i.e. 2 for the second column)</td>
<td></td>
</tr>
</tbody>
</table>

SELECT COUNT(employee_id), department from EMPLOYEES where first_name = "" group by department order by 2 |

16 See explanation | A check to see if a basic file open, read, close can be ordered. Simple process, but some may not be disciplined in the close process. Lines 4 and 5 will check some data structure understanding and cannot be completed until after the file has been read. |

f = open("ads.csv")
data=f.read()
f.close()

FirstTenChars = data[0:29]
print(FirstTenChars) |

17 C A basic test of statistical knowledge - specifically how to calculate the mean of a number. Put into a Python context adds no difficulty but does test some understanding of algorithmic steps. | 5.1, 5.2, 5.3, 5.4, 5.5 |
<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>Explanation / Rationale</th>
<th>Syllabus Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>D</td>
<td>All of the functions are valid but only summary will return more than one average as well as the quartiles.</td>
<td>5.1, 5.2, 5.3, 5.4, 5.5</td>
</tr>
<tr>
<td>19</td>
<td>A</td>
<td>This checks a basic function of data visualisation in R. Head is a common function, some may not know that the default amount of rows returned is 6, but a process of elimination will lead them there if they understand what's wrong with the other options.</td>
<td>5.1, 5.2, 5.3, 5.4, 5.5</td>
</tr>
<tr>
<td>20</td>
<td>See explanation</td>
<td>A more complex set of steps as it involves slightly more advanced functions. They will not need to remember syntax but should understand the arguments in the ts function as they are the only option given.</td>
<td>5.1, 5.2, 5.3, 5.4, 5.5</td>
</tr>
</tbody>
</table>

BirthsByYear <- read.csv("births.txt")
birthstimeseries <- ts(BirthsByYear)
plot.ts(birthstimeseries)
birthstimeseries <- ts(BirthsByYear, frequency = 12, start = 1999)
monthplot(birthstimeseries)