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 40 questions in total across:

- 20 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.
- 4 scenario-driven situational judgement assessments 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 your answer sheet.

Pass mark is 26/40
Distinction is 34/40
Time allowed: 90 minutes

This is a United Kingdom government regulated qualification which is administered and approved by one or more of the following: Ofqual, Qualifications Wales, CCEA Regulation or SQA.

1 Your organisation needs to collate data from different internal departments, as well as data originating from external stakeholders. Which ONE of the following options would be necessary before this dispersed data can be used effectively?

A It will need to be analysed and summarised.
B It will need to be cleaned and standardised.
C It will need to be translated and shared.
D It will need to be expanded and visualised.

2 Under which of the following conditions would there MOST LIKELY be a potential issue when working with personal financial information?

A The data has been stored within a customer database which can be accessed by internal staff.
B The data has been stored in a downloads folder accessed by a member of internal staff.
C The data has been stored on a well-known cloud storage service.
D The data has been stored on a remote company data centre which can be backed up remotely.

3 Which of the following data types could BEST be described as personally identifiable information (PII)?

A Your customers' internal ID numbers.
B How customers' purchase volume has changed over time.
C Shipping addresses for your customers' most recent orders.
D Purchases your customers have made.

4 You are currently gathering data relating to coastal erosion, in order to predict the measurement of the erosion in another 20 years' time. It includes a series of measurements that have been recorded over a period of 50 years. Which of the following types of data would you MOST LIKELY use in your analysis?

Choose TWO of the following options.

A Continuous.
B Descriptive.
C Structured.
D Nominal.

5 A customer is using an online shopping app and has recently updated their personal details via their online account. It is found later that the organisation still only has their original details when the product the customer ordered was delivered to the wrong address.

At which point in the data lifecycle is this issue MOST LIKELY to have arisen?

A Creation.
B Storage.
C Use.
D Deletion.

6 Which of the following data structures includes the use of a parent node?

A Graph.
B List.
C Tree.
D Array.

7 Which of the following best describes why a relational database is a structured data source?

A Data is tagged with an attribute.
B Data is organised sequentially.
C Data is scalable.
D Data has a pre-defined format.

8 Which of the following unstructured data formats would require the MOST amount of pre-processing before the data can be used effectively?

Select the TWO MOST LIKELY options.

A Audio.
B Word processed files.
C Video.
D Social media feeds.

9 You want to gain insight into the influence your customers have on brand visibility. You have a structured data source in the form of a CRM, as well as unstructured data from social media feeds.

What would be the MAIN benefit of using this unstructured data alongside the structured data source?

A You could identify how active your customers are on social media.
B You could identify how many times your brand's name is mentioned.
C You could identify the number of followers your customers have.
D You could identify how many times your main competitor's name is mentioned.

10 Which of the following options would MOST LIKELY be detrimental to data analysis?

A Incomplete data.
B Out-of-date data.
C Duplicate data.
D Unverified data.

11 You have been tasked to produce a monthly report on sales from the previous month. What sort of analytics would you use?

A Decision analytics.
B Descriptive analytics.
C Predictive analytics.
D Prescriptive analytics.

12 An organisation's Data Protection Officer has been asked to carry out a GDPR-compliant right to erasure request. Which of the following will they will need to know?

Select all that apply.

A Who the information relates to.
B Why the information has been stored.
C Where the information is stored within the organisation's systems.
D When the information was captured.

13 Match each of the following GDPR-related roles to the appropriate description of their responsibilities.

GDPR-related roles	*Description*
1. Controller	[ ] Responsible for monitoring GDPR compliance.
2. Processor	[ ] Processes personal data on behalf of someone else.
3. Data Protection Officer	[ ] Responsible for overseeing data protection strategy and its implementation.
4. Supervisory authority	[ ] Determines the purposes for and means of the processing personal data.

14 A business wants to run a data analysis project to assess the financial viability of ongoing projects. Sort the following options into **business requirements** and **technical requirements** by deleting the option as appropriate.

A Produce a finance report at the end of each month. **Business/technical**
B Restrict data access to privileged users. **Business/technical**
C Give a viability score to each project. **Business/technical**
D Format the data as a csv. **Business/technical**

15 In terms of requirements elicitation, which of the following options would MOST LIKELY be used in order to monitor a current process?

A Recounting.
B Observation.
C Enacting.
D Technical testing.

16 You want to create a data model which describes the technical requirements of a data analysis project. The intended audience will be non-technical company directors. Select the MOST appropriate data model from the following options.

A Conceptual data model.
B Logical data model.
C Physical data model.
D Technical data model.

17 Which of the following could be described as tacit knowledge?

Choose all that apply.

A Procedures.
B Intuition.
C Experience.
D Databases.

18 Which of the following should define what data is collected and stored in an organisation?

A Policies.
B Standards.
C Storage space.
D Processing availability.

19 Which of the following data architecture functions would support business intelligence activities on historical data?

Choose ONE option.

A Data modelling.
B Data integration.
C Data warehousing.
D Data migration.

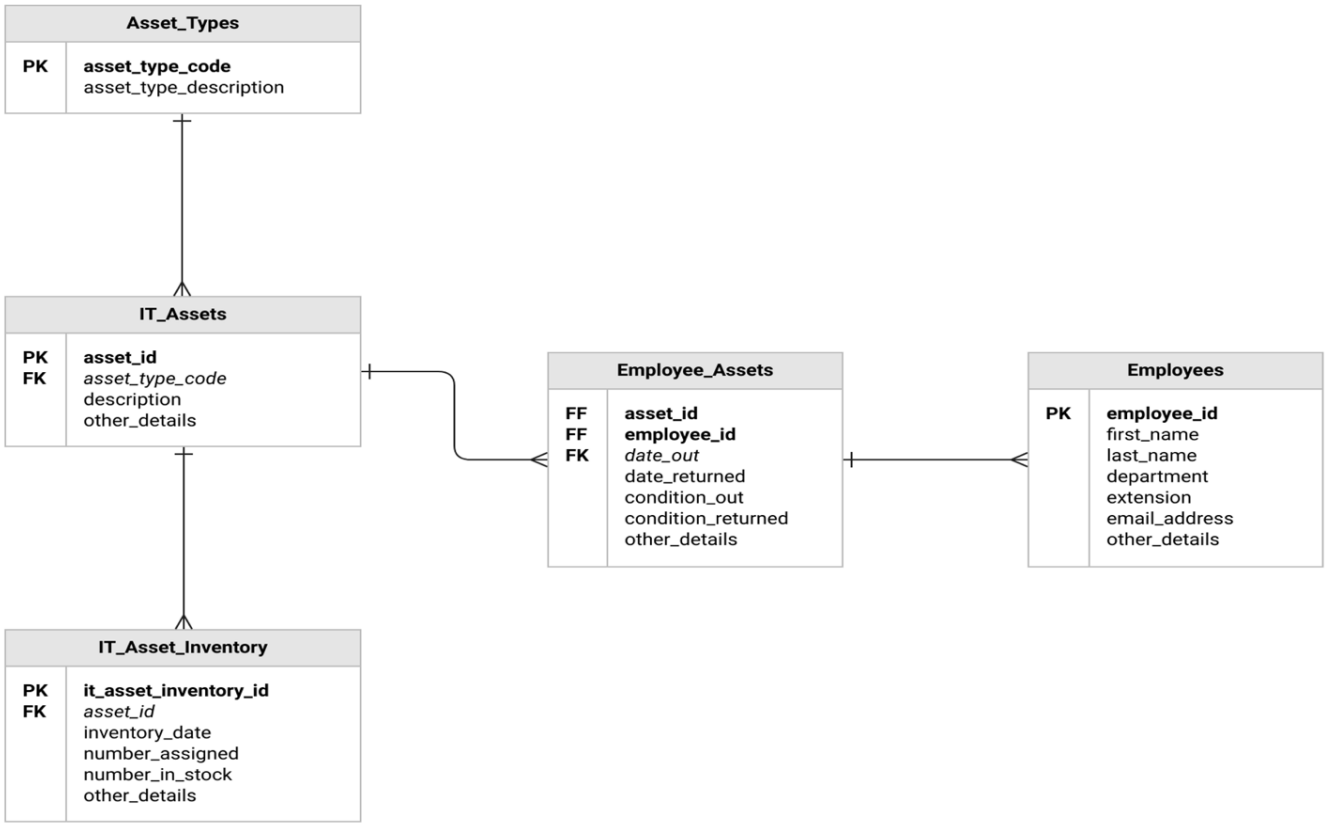
20 Which of the following are the key challenges in dealing with Big Data?

Choose all that apply.

A Volume of data.
B Velocity of data.
C Variety of data.
D Verification of data.

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.



21 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

22 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

23 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

24 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

25 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.

26 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.

```
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:

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

```
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

28 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)

29 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]

30 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)

Scenario 3 – ERD Normalisation

You have been asked to lead a project working on the data used by a retail business. They currently capture all the key sales information in a spreadsheet but would like to migrate this to a relational database. The structure of the spreadsheet is shown below with five rows of data.

Salesperson number	Salesperson Name	Customer Number	Customer Name	Product bought	Purchase date	Sale amount
3212	Bloggs	54	Adams	Table lamp	12/03/2021	54.65
		23	Blight	Chair	12/03/2021	345.99
4322	Smith	56	Montagne	Table	20-Mar	564.99
		43	Wayne	Sofa	19-Mar	657.89
431	Jones	21	Clarke	Chair	19-Mar	345.99
		45	Delta	Table lamp	20-Mar	54.65
		67	Levy	Footstool	12/03/2021	56.79
877	Tremlett	56	Montagne	Pillow	18-Mar	12.35
		43	Fargo	Cushion	20-Mar	7.69
877	Farquar	11	Plymo	Sofa	12/03/2021	657.89
		89	James	Table lamp	25-Mar	54.65
		10	Jones	Chair	12/03/2021	345.99

31 Which of these outcomes would you expect when normalising this data to first normal form?

A No change.
B Repeat the Salesperson information and have 12 separate rows of data.
C Delete the Customer Number column.
D Creation of a two or more separate tables.
E Sales amounts would be aggregated.

32 Which tables could be created in a first normalised form?

A Discount table.
B Purchase table.
C Customer table.
D Shop location table.
E Salesperson table.

33 If you created a purchase table storing information about the product, price and date of purchase, what other field should be added?

A Salesperson number.
B Customer number.
C A new primary key – PurchaseID.
D Currency exchange rate.
E Table creation date.

34 Creating the normalised form of this data would be categorised as which form of model?

A Physical model.
B Business model.
C Conceptual model.
D Database model.
E Technical model.

35 Creating a database for the retail data will required discussions with stakeholders to clarify requirements. Which of the following techniques would be appropriate for this scenario?

A Prototype.
B Interviews.
C Process modelling.
D Financial costing.
E Return on investment.

Scenario 4 – Data Modelling

You have been asked to show senior managers what a statistical analysis of data could be used for in the business. You have decided to create a predictive model to forecast sales of products within the company's stores.

36 In planning your approach to the predictive model, you have decided to create a project plan with the key steps required. Order the given steps in the correct order (earliest to latest). Indicate your chosen order by writing the letters A-E in the spaces below.

A Cleanse data.
B Create problem hypothesis.
C Analyse data.
D Collect data.
E Document results.

__ __ __ __ __

37 You feel that weather has a lot to do with sales in stores so you would like to create a model to show the impact that weather has. You believe that sunny weather increases total sales. Which of the following statements would form a suitable null hypothesis for this model?

A H0 - the amount of daily sunshine increases the total daily sales.
B Ha - the amount of daily sunshine increases the total daily sales.
C H0 - the amount of daily sunshine does not impact the total daily sales.
D Ha - the amount of daily sunshine does not impact the total daily sales

38 Having seen that weather does not appear to impact sales in stores, you have decided to forecast sales using a linear regression model. Having defined the model, you are looking to now train the model. What would be an appropriate size subset of your data to use for this training?

A 0%
B 20%
C 30%
D 50%
E 70%

39 How much data should you set aside for your testing and validation data sets?

A 30%
B 50%
C 70%
D 80%
E 100%

40 You are looking to present the results of your linear regression model to senior stakeholders. Which visualisation would be most appropriate for a linear regression forecast?

A Bar chart.
B Histogram.
C Gantt chart.
D Heat map.
E Scatter chart.

End of Paper

Question	Answer	Explanation / Rationale	Syllabus Section
1	**B**	Dispersed data from multiple different sources will need to be cleaned and standardised before it is usable.	1.1
2	**B**	There is a potential security issue here in that a physical copy of sensitive information is being held on an individual's device rather than in once secure location, which creates the risk to the information should the device be lost, stolen or compromised.	1.2
3	**C**	No other options can be used to directly identify individual people.	1.2
4	**A and C**	Continuous data is any value that can take on any value within a certain range, such as a measurement of temperature or distance. This type of numeric data would also most likely to be structured.	1.3
5	**B**	Even though the customer has updated their details via the website, the issues appears to be with how the data is being stored i.e. the new details have not been stored in place of the old ones, or the new data has gone to the wrong place.	1.4
6	**C**	A tree would be used to group data in a hierarchical order using parent, child and leaf nodes.	2.1
7	**D**	Option A relates to XML. Option B - sequentially organised data is structured, but relational databases do not organise data in this way. Option C does refer to relational databases, but this is not what makes it structured.	2.2
8	**A and C**	Unstructured data from audio/video sources would be processed differently to text, as the user would need to convert the audio to text first.	2.4
9	**B**	Businesses can use unstructured data sources to enrich structured data when used in a specific and focused way, i.e. they should have a goal in mind.	2.5
10	**B**	Duplicate records can be ignored when processing; incomplete data can still add value; unverified data can't be trusted as much as verified data can be, but is not necessarily incorrect.	3.1
11	**B**	A historic view of data without any outcomes would be classed as descriptive.	3.2
12	**A and C**	The organisation would need to know who the data subject is and where their information is stored, in order to locate and delete it.	3.3
13	**See rationale**	1. Determines the purposes for and means of the processing personal data. 2. Processes personal data on behalf of someone else.	3.4

Question	Answer	Explanation / Rationale	Syllabus Section
		3. Responsible for overseeing data protection strategy and its implementation. 4. Responsible for monitoring GDPR compliance.	
14	**A-C Business D Technical**	The format of the data is a technical requirement; while options A-C have a technical element to them, they are business requirements.	3.5
15	**B**	This technique is an effective means for monitoring a current or ongoing process, for example, deciphering how a user performs their role by assessing their work environment.	3.6
16	**A**	Physical and logical data models would be too technical for the intended audience; conceptual data model identifies business concepts and is used to document business from a data perspective - more about actual business data rather than database design.	3.7
17	**B and C**	Unwritten and hard to communicate knowledge of a business - not usually documented in any form but resides with experts	3.7
18	**A and B**	Data architecture should be defined by policies, rules and standards rather than available resource or technical constraints.	4.1
19	**C**	Data warehousing is the only one of the four options which relates to historical data	4.2
20	**A, B and C**	Acknowledging the 3 V's of describing Big Data challenges - there are other accepted descriptions but verification is not a relevant term.	4.3
21	**E**	The answer field implies an 'id' based on manager which should lead through thought process to unique identifier of manager.	4.4, 4.6
22	**D**	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.	4.4, 4.6
23	**A and B**	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'.	4.7
24	**C**	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.	4.7

Question	Answer	Explanation / Rationale	Syllabus Section
25	**See explanation**	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) SELECT **COUNT**(employee_id), department from **EMPLOYEES** where first_name = "" group by **department** order by **2**	4.7
26	**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)	5.1, 5.2, 5.3, 5.4, 5.5
27	**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
28	**D**	All of the functions are valid but only summary will return more than one average as well as the quartiles.	5.1, 5.2, 5.3, 5.4, 5.5
29	**A**	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 whats wrong with the other options.	5.1, 5.2, 5.3, 5.4, 5.5
30	**See explanation**	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. BirthsByYear <- read.csv("births.txt") birthstimeseries <- ts(BirthsByYear) plot.ts(birthstimeseries) birthstimeseries <- ts(BirthsByYear, frequency = 12, start = 1999)	5.1, 5.2, 5.3, 5.4, 5.5

Question	Answer	Explanation / Rationale	Syllabus Section
		monthplot(birthstimeseries)	
31	**B and D**	Each record has to have a primary key (and purchases does not currently have one). Each record cannot have repeating groups of attributes therefore Salesperson needs to be separated into individual records for each customer/purchase.	2.2, 3.8, 4.4, 4.6
32	**B, C and E**	The Salesperson and Customer data are clear sets of data with purchase information (date and amount) also a valid choice to store. There is no indication of discount or location data so these tables would not be an outcome from this process.	2.2, 3.8, 4.4, 4.6
33	**C**	The new table should always have a unique field for the primary key. There is no obvious choice for this in the current fields so a new ID field should be created.	2.2, 3.8, 4.4, 4.6
34	**C**	The conceptual model is to establish the entities, their attributes, and their relationships. The logical data model defines the structure of the data elements and set the relationships between them. The physical Data Model describes the database-specific implementation of the data model.	2.2, 3.8, 4.4, 4.6
35	**A, B and C**	As a conceptual model, this is aimed at process requirements elicitation. The financial measures are not relevant at this stage but the other options are all potential opportunities to gather requirements.	3.7
36	**B, D, A, C, E**	Some of the typcial steps of data analysis in the correct order (as per syllabus).	6.1
37	**C**	Null hypothesis is indicated by H0 and would state that there is not enough statistical evidence to show that sunshine affects product sales.	6.2
38	**E**	Normal practice for size of a training data set is at least 70% (sometimes increasing but 70% option is the only realistic choice in this list).	6.4
39	**A**	The remaining data after the training data set is taken out would be used for testing/validation.	6.5
40	**E**	As a linear model, the most appropriate visualisation should show the data points and the linear trend line i.e. Scatter Chart.	6.6