BCS Level 4 Data Analyst Module

Sample Paper A

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

Sample paper consisting of 40 questions in total, including:

- 20 multiple choice knowledge questions – 1 mark awarded for each question.
- 4 scenario-driven, situational judgement assessments, each with five questions designed to test knowledge, skills and behaviours. These 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 D (**or **E).**

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 mark 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 Which **one** of the following data types could **best** be described as Personally Identifiable Information (PII)?

A Customer ID numbers used in internal systems.
B Data showing how customers' purchasing patterns have changed over time.
C Shipping addresses for customers' most recent orders.
D Aggregate data on product popularity amongst customers.

2 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 **one** of the following types of data would you **most likely** use in your analysis?

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

3 Which **one** of the following data structures includes the use of a parent node?

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

4 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 customer relationship management (CRM) system, 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.

5 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.

6 You want to create a data model that 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.

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

A Architectural policies.
B Network infrastructure.
C Storage space.
D Processing availability.

8 Which **one** of the following data architecture functions would support business intelligence activities on historical data?

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

9 The rate at which data is generated, collected, processed and analysed describes which challenge associated with big data?

A Volume.
B Velocity.
C Variety.
D Veracity.

10 When talking about big data, which **one** of the following could you reasonably expect to see in your original datasets?

A Predicted values.
B Advanced statistical methods.
C Unstructured content.
D Machine Learning model.

11 Which **one** of following would you expect to feature in an entity relationship diagram (ERD)?

A Attributes.
B Timings.
C Object states.
D Methods.

12 Which **one** of the following is an advantage of a relational database over a NoSQL database?

A Easily take advantage of parallel processing.
B Speed of transactions for low volumes of data.
C Flexible schema that can be easily changed.
D Attributes can be added anytime.

13 Which stage of ETL would cleanse data?

A Encrypt.
B Transform.
C Extract.
D Load.

14 Which **one** of the following operations would you expect to the happen in the Extract phase of ETL?

A Deriving calculated fields.
B Applying business rules.
C Data cleansing.
D Staging data from source systems.

15 Which **one** of the following has a key purpose of being able to create "a highly intuitive, drag-and-drop interface for building visualisations, reports, and dashboards"?

A R.
B Python.
C PowerBI.
D Excel.

16 Which **one** of the following is a drawback when using the mean average of a set of data?

A Extreme values can heavily influence the result.
B It is difficult to find tools that calculate the value.
C It will not work well with accurate figures.
D It does not always provide a single value as the answer.

17 Which **one** of the following is a drawback when using the mode average of a set of data?

A Extreme values can heavily influence the result.
B It may not provide a single value as the answer.
C It may be necessary to interpolate the values.
D It is not available as a function in spreadsheet software.

18 Which **one** of the following is a drawback when using the median average of a set of data?

A The result may not appear in the original data set.
B It does not always provide a single value as the answer.
C Extreme values can heavily influence the result.
D It is not available as a function in Excel.

19 When considering strategies to improve data analysis modelling, which **one** of the following methods is **most** effective?

A Randomising values and re-running the model.
B Reducing the size of the testing set.
C Adding more data to the testing set.
D Adding more data to the training set.

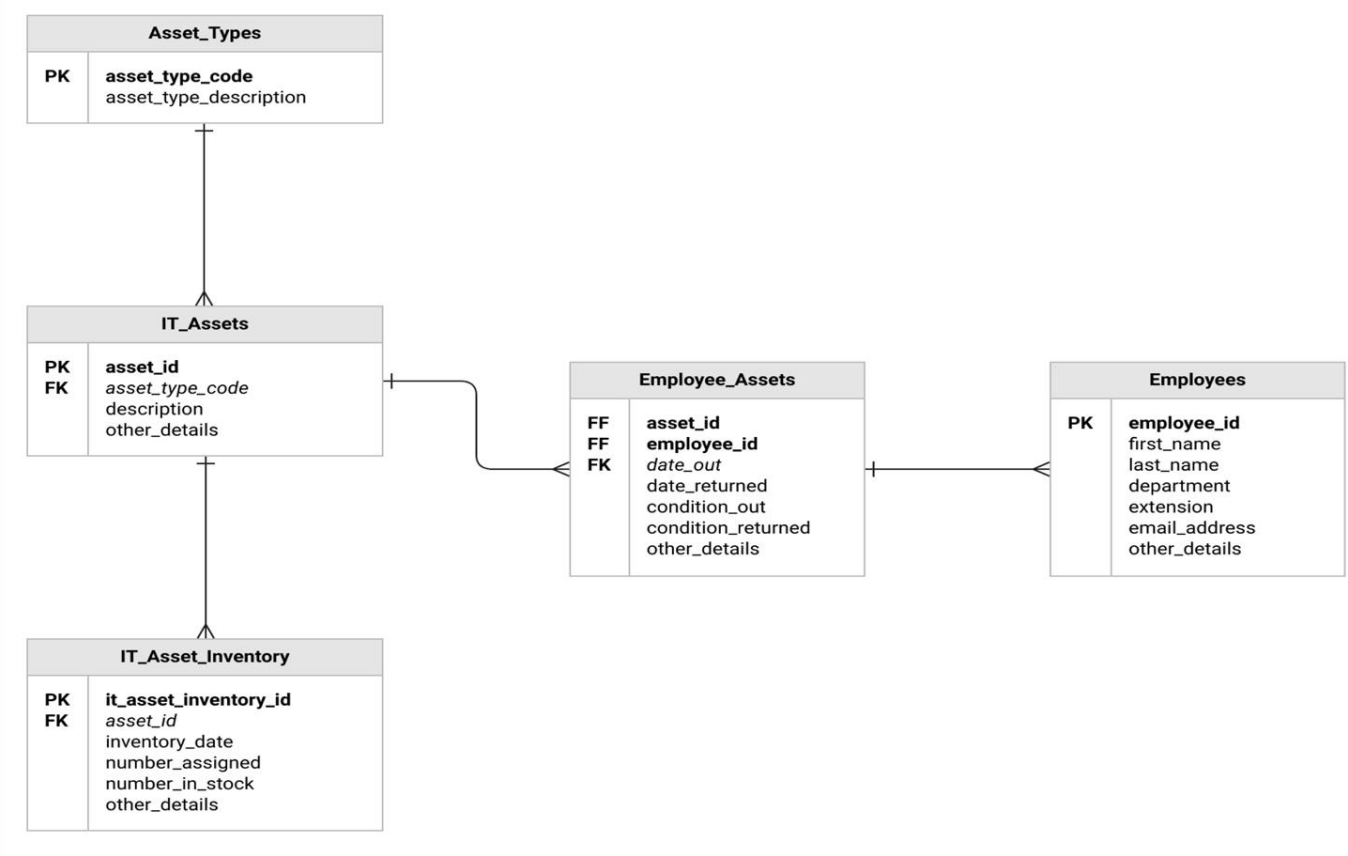
20 When visualising data for stakeholders, which **one** of the following is the key factor to consider?

A The audience's understanding of the coding used.
B The accessibility of the presented data and ease of understanding.
C The qualifications you hold in data analytics.
D The time and date that the visualisations are presented.

Scenario 1: Database design and SQL

This scenario relates to questions 21 - 25.

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 would be a suitable primary key for the 'Managers' table?

A employee_id.
B asset_id.
C asset_type_code.
D first_name and last_name.

22 You have been tasked with creating a data dictionary for this database.

Which **one** of the following should be listed in a data dictionary?

A table records.
B typical errors.
C variable antonyms.
D field size.

23 A key report measure on the scorecard is the number of assets returned in a month, and their condition when returned.

Which **two** 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.

What database functionality ensures that this happens?

A Delete query.
B Update query.
C Cascading delete.
D Index.

25 You require a report that shows 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.

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

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

This scenario relates to questions 26 - 30

You are working as a data analyst in an 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 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.

You can use the space below to provide your answer, unless using a separate answer sheet.

f.close
print(FirstTenChars)
data=f.read()
f = open(“ITOutages.csv”)
FirstTenChars = data[0:10]

27 Which **one** of the following is the missing line of code in the Python programme below 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=Total**2.

28 The IT department are also considering using the language R.

Which **one** 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)

29 Which **one** of the following R commands would be used 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]

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 then plot a second graph that starts in 1999.

You can use the space below to provide your answer, unless using a separate answer sheet.

employeetimeseries <- ts(EmployeesByYear, frequency = 12, start = 1999)
monthplot(employeetimeseries)
plot.ts(employeetimeseries)
EmployeesByYear <- read.csv("ITemployees.txt")
employeetimeseries <- ts(EmployeesByYear)

Scenario 3 – ERD Normalisation

This scenario relates to questions 31 – 35.

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 **two** of the following outcomes would you expect when normalising this data to first normal form?

A There would be no change.

B The Salesperson information would be repeated, creating 12 separate rows of data.

C The Customer Number column would be deleted.

D Two or more separate tables would be created.

E Sales amounts would be aggregated.

32 Which **three** of the following tables could be created in a normalised form of the data?

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

33 If you created a table storing information about salespeople, assuming the salesperson number is unique, what type of key would this make?

A Foreign key.
B Primary key.
C Secondary key.
D Composite key.

34 When producing models as part of your database design, 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.

35 To work on your data analytics project using this database, you will need specify what your hypothesis is. You have been asked to investigate the impact of the purchase date on sales.

Which **one** of the following statements would be an appropriate null hypothesis?

A The later the purchase date, the more money the customer spends.
B The purchase date has no impact on the number of sales.
C The earlier the purchase date, the more money the customer spends.
D The amount of spending has no impact on the purchase data.

Scenario 4 – Data Modelling

This scenario relates to questions 36 – 40.

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 your preparation of the data, you have decided to import it into Python for manipulation. Your dataset contains different types of data and you are going to rely on the 'panda' library.

In your initial analysis of two columns containing numeric data, you have checked the two columns against each other. If the first column (col1) contained "1" and the second column (col2) contained "0", what would be returned from col1 AND col2?

A 0.
B 1.
C 10.
D 01.

37 You would like to create a model to show the impact that weather has on sales in stores. You believe that sunny weather increases the total sales.

Which **one** 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 30%.
C 50%.
D 70%.

39 One of the managers claims that there is a correlation between the number of people in the town centre and the number of sales. Whilst this sounds logical, your model shows a correlation coefficient of 0.4.

How would you interpret this coefficient result?

A Sales are strongly correlated to the number of people in the town centre.
B Sales are negatively correlated to the number of people in the town centre.
C Sales are inversely correlated to the number of people in the town centre.
D Sales are moderately correlated to the number of people in the town centre.

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 Scatter chart.

End of paper

BCS Level 4 Data Analyst Sample Paper
Answer Key and Rationale

Question	Answer	Explanation / Rationale	Syllabus Section
1	**C**	Personally Identifiable Information refers to any data that could potentially identify a specific individual. No other options can be used to directly identify individual people.	3.4
2	**A**	Continuous data can take any value within a certain range, such as a measurement of temperature or distance.	1.3
3	**C**	A tree would be used to group data in a hierarchical order using parent, child and leaf nodes.	2.1
4	**B**	Businesses can use unstructured data sources to enrich structured data when used in a specific and focused way. Unstructured social media data can provide valuable insights into brand visibility, complementing the structured data from the CRM.	2.5
5	**B**	A historic view of data without any outcomes would be classed as descriptive analytics.	3.2
6	**A**	A conceptual data model is a high-level representation of organisational data, designed to be easily understood by stakeholders, including non-technical users. Its primary purpose is to illustrate the entities (or concepts) within a system and the relationships between those entities, without delving into the technical details. Physical and logical data models would be too technical for the intended audience.	3.7
7	**A**	Data architecture should be defined by policies, rules and standards, rather than available resource or technical constraints.	4.1
8	**C**	A data warehouse is designed to store and manage large volumes of historical data, making it an ideal solution for supporting business intelligence activities, such as reporting, data analysis, and decision-making.	4.2
9	**B**	Velocity refers to the speed at which data flows into systems, and the need for systems to handle and process this data quickly and efficiently.	4.3

Question	Answer	Explanation / Rationale	Syllabus Section
10	**C**	A key feature of big data is the likely inclusion of unstructured data. All of the other options are methods/techniques that you would apply to the dataset when modelling, rather than being a feature of the original data.	4.3
11	**A**	Attributes are a key part of an entity and would therefore be shown on an ERD.	4.4
12	**B**	Relational databases are designed to work in a formalised manner on known data structures. This makes them far more appropriate for high transactional situations. Relational databases do not easily change, so flexible schemas and changing attributes are not advantages. Relational databases also cannot be 'split' in the same way as NoSQL, so are less suited to parallel processing.	4.5
13	**B**	Transform deals with all processing of data after it has been extracted and before loading clean/calculated data into the destination system	5.1
14	**D**	Extracted data is often temporarily stored in a staging area before transformation. This staging area allows for further processing and ensures that the extraction process does not impact the performance of source systems. The remaining options are all part of the transform stage.	5.1
15	**C**	PowerBI is the only option that provides a drag and drop interface for dashboards. Excel has some guided functionality, but this is limited and not a key purpose of the software. R and Python have GUI IDEs available, but these are for programming and coding rather than drag and drop creation of dashboards.	5.3
16	**A**	Very high or very low values will change the mean significantly. It is a standard function in Excel (which defaults to mean). It does not get complicated by accuracy of figures (which mode does). It will always provide a single answer (which mode will not).	5.4

Question	Answer	Explanation / Rationale	Syllabus Section
17	**B**	Datasets can have more than one value that appears the most frequently. A - refers to mean rather than mode. C - Interpolation maybe required when using median (not mode). D – it is a less well-known function in Excel.	5.4
18	**A**	Median values can be halfway between two values in the dataset and therefore may not exist as a point of data. This is a deeper understanding of the value of different summary statistics with different types of dataset. It will provide a single value (unlike mode) and it is not unduly influenced by extreme values (unlike mean).	5.4
19	**D**	Greater amounts of data in the training set usually provides a greater basis for modelling.	6.4
20	**B**	Communicating information requires thought about how it will be received and in what format. The code/program used will not be a prime consideration - this is only valid when discussing the modelling, not visualising results. Your qualification level is not a discussion for the visualisation of results (this maybe a factor at the outset of a project). The timing of the presentation is a consideration for the audience enthusiasm but will have no effect on the visualisation decisions.	6.6
21	**A**	Logical reasoning should lead to assumption that a manager would also be an employee. It would therefore be appropriate to use an employee ID as a unique identifier. The other options are unrelated to an employee or manager (asset_id and asset_type_code) and firstname with lastname is still not guaranteed to be unique (and this can also be derived from the current structure not using this as an ID in the employee table).	4.4, 4.6

Question	Answer	Explanation / Rationale	Syllabus Section
22	**D**	Data dictionaries contain information about variables. A key element is the size of the field that should be used for each variable. Antonyms is a distractor and not a valid term in any context. Allowed values would commonly be in the dictionary but typical errors would not. Table records is a physical database design concern rather than in a list of variable terms - data dictionary.	4.4
23	**A and B**	The Count aggregation will provide the total number of assets returned in a month. The Group By aggregation is used to group rows that have the same values i.e. condition.	4.7
24	**C**	Cascading delete is a database feature that automatically deletes related records in child tables when a record in the parent table is deleted. This ensures data integrity by removing associated data across tables.	4.7
25	**See explanation**	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 (department). SELECT COUNT(employee_id_, department FROM EMPLOYEES WHERE first_name = “” GROUP BY department ORDER BY 2	4.7
26	**See explanation**	f = open("ITоutages.csv") data=f.read() f.close() FirstTenChars = data[0:10] print(FirstTenChars)	5.3

Question	Answer	Explanation / Rationale	Syllabus Section
27	**C**	This line calculates the mean by dividing the total sum of the numbers by the number of elements in the list, which is 5 in this case.	5.3, 5.4
28	**D**	The summary() function in R provides a summary of the statistical properties of the dataset, including measures such as mean, median, quartiles, minimum, and maximum values.	5.3, 5.4
29	**A**	The head() function in R is used to view the first few rows of a dataset. By default, it displays the first six rows of the dataset. head(dataset, 5) would show the first five rows, not the first six rows. tail(dataset, 6) would show the last six rows, not the first six rows. dataset[6] would only display the sixth row of the dataset, not the first six rows.	5.3, 5.4
30	**See explanation**	EmployeesByYear <- read.csv("ITemployees.txt") employeetimeseries <- ts(EmployeesByYear) plot.ts(employeetimeseries) employeetimeseries <- ts(EmployeesByYear, frequency = 12, start = 1999) monthplot(employeetimeseries)	5.2, 5.3
31	**B and D**	Each record must 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. This will turn 5 records into 12 separate rows. There is also information about other entities which means normalising will lead to more than one table being created. Question is not asking for the results from normalisation - it is looking to see understanding in the concepts (which would lead to individual rows and separate table).	4.4, 4.6

Question	Answer	Explanation / Rationale	Syllabus Section
32	**B, C and E**	Normalising will likely create tables for the entities shown - the learner does not need to complete the process, just understand what shape the outcomes will take. As there is no indication of discount or location data in the dataset, these tables would not be an outcome from a normalisation process.	4.4, 4.6
33	**B**	This question checks that learners understand the difference between composite keys and foreign/primary keys. Primary keys are unique identifiers in a table.	4.4, 4.6
34	**C**	The conceptual model is to establish the entities, their attributes, and their relationships. The physical data model describes the database-specific implementation of the data model. The physical model would show the data types and database table design. The business and database model are not relevant terms in this topic area.	4.4, 4.6
35	**B**	The null hypothesis states **no** relationship between the variables selected.	6.2
36	**A**	This is a check of the boolean operators. The use of AND on a 1 and 0 will return 0.	5.4
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	**D**	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). Using 50% or less would not be appropriate for a training set.	6.4
39	**D**	Correlation coefficients range from -1 to 1. Learners should understand the key regions of values with negatives showing negative and inverse relationships. The figure would have to be closer to 1 than 0 to show a strong correlation - therefore the correct answer is moderate.	5.4

Question	Answer	Explanation / Rationale	Syllabus Section
40	**D**	As a linear model, the most appropriate visualisation should show the data points and the linear trend line i.e. scatter chart.	6.6