Assessor Guide

BSBXBD403

Analyse big data

Assessment 1 of 4

Quiz

Version 1

Assessment Instructions

Task overview

This assessment task contains 20 quiz questions which consist of the following:

- One (1) x Ordering
- Fourteen (14) x Matching
- One (1) x True or False
- Four (4) x Multiple choice

Read each question carefully before selecting your responses from the options provided.

Assessment Information

Submission

You are entitled to three (3) attempts to complete this assessment satisfactorily. Incomplete assessments will not be marked and will count as one of your three attempts.

All questions must be responded to correctly to be assessed as satisfactory for this assessment.

Answers must be typed into the space provided and submitted electronically via the LMS. Hand-written assessments will not be accepted unless previously arranged with your assessor.

Reasonable adjustment

Students may request a reasonable adjustment for assessment tasks.

Reasonable adjustment usually involves varying:

- the processes for conducting the assessment (e.g. allowing additional time)
- the evidence gathering techniques (e.g. oral rather than written questioning, use of a scribe, modifications to equipment)

However, the evidence collected must allow the student to demonstrate all requirements of the unit.

Refer to the Student Handbook or contact your Trainer for further information.

Question 1

Match each big data analysis scenario with the relevant legislative requirement from the list of options (A to F) provided.

Legislative requirements:

A. Privacy Act 1988 *(Long URL: https://www.legislation.gov.au/Details/C2022C00135]*
B. Data-matching program Act 1990 *(Long URL: https://www.legislation.gov.au/Details/C2020C00371)*
C. Competition and Consumer Act 2010 (Section 56BA) *(Long URL: https://www.legislation.gov.au/Details/C2021C00528)*
D. General Data Protection Regulation (GDPR) *(Long URL: https://www.oaic.gov.au/privacy/guidance-and-advice/australian-entities-and-the-eu-general-data-protection-regulation)*
E. Data Availability and Transparency Act 2022 *(Long URL: https://www.legislation.gov.au/Details/C2022A00011)*
F. Privacy and Personal Information Protection Act 1998 *(Long URL: https://legislation.nsw.gov.au/view/whole/html/inforce/current/act-1998-133)*

Table 1 - Quiz Q1: Legislative requirements for big data analysis

Big data analysis scenario:	Answer	Legislative requirements
The analysis of a dataset that includes customer information obtained from several NSW public sector agencies.	Option F	Option A
As part of the analysis the records from one dataset was linked with another to find records that have related information. As a result of this, the dataset is found to contain individual tax file number details.	Option B	Option B
An Australian Government dataset is shared with an organisation for the purpose of enabling statistical analysis for research and development according to a data sharing agreement.	Option E	Option C
A dataset that contains information about individuals from their online activities (e.g. their tastes and preferences) from online purchases and transaction history data is analysed. However, the gathered information may be incorrect.	Option A	Option D
A dataset was obtained from a third party for analysis. During analysis, the dataset was found to contain information from an identifiable small business with an annual turnover of $3 million or less. There's no evidence of a valid Consumer Data Right (CDR) request or consent made to the identified small business.	Option C	Option E
A dataset for analysis was collected from an Australian business that engages with overseas businesses and contractors in the European Union (EU) offering goods and services in the EU. The dataset contains information on EU residents. The purpose of the analysis is to forecast future sales and services of goods.	Option D	Option F

Question 2

Match each data analysis activity example that relates to analysing transactional and non-transactional big data with the correct analytical techniques and tools from the list of options (A to E) provided.

Analytical techniques and tools:

A. Data mining
B. Ad hoc queries
C. Operational business intelligence (OBI)
D. Real-time business intelligence (RTBI)
E. Text analysis

Table 2 - Quiz Q2: Analytical techniques and tools

Data analysis activity examples	Answer	Analytical techniques and tools
The analysis involves identifying customer sentiment towards the organisation's products and brand based on customer product reviews.	Option E	Option A
Analysing loyalty card data generated from a supermarket store to decide when to offer customers coupons targeted to their buying habits and decide when to put items on sale or when to sell them at full price.	Option A	Option B
An analyst is asked to investigate daily average users and has found that one day it is 5% lower than the previous day. The analysis involves investigating further into why the number of users has declined on that specific day.	Option B	Option C
The analysis of business processes of a product warehouse to offer a solution that creates automated alerts and workflows that provides non-managerial staff with information such as inventory and product re-order levels.	Option C	Option D
Analysing a logistics business workflow to provide a solution to track over 50,000 products at different store locations. The purpose was to gain insights into optimising product delivery routes, loading and unloading times based on dynamically changing demands.	Option D	Option E

Question 3

Match each description with the associated organisational policy and/or procedure that relates to analysing big data from the list of options (A to F) provided.

Policies and procedures:

A. Data analysis procedure
B. Data categorisation procedure for analysis
C. Data source identification policies and procedures
D. Procedure for integrating big data sources
E. Procedure for combining external big data sources
F. Policies and procedures for reporting on big data analysis

Table 3 Quiz Q3: Organisational policies and procedures

Description	Answer	Policies and procedures
Involves a set of procedures for establishing and confirming the types and groupings of the data to be applied in the analysis.	Option B	Option A

Description	Answer	Policies and procedures
Provides guidelines on how to combine big data obtained from sources outside of the organisation (e.g. social media), with in-house big data.	Option E	Option B
Provides guidelines on how to combine datasets from different sources that may have different formats (e.g. structured, semi-structured and unstructured).	Option D	Option C
Provides guidelines on how to evaluate and investigate the big dataset to find patterns, trends, and relationships to make sense of the data and to identify business insights.	Option A	Option D
Provides guidelines on how to recognise where a dataset is sourced from and understanding how to use the dataset appropriately whilst complying with any legislative requirements that may apply.	Option C	Option E
Provides best practices and guidelines to ensure that the analysis of the big dataset is presented clearly and consistently using standard representations (e.g charts, graphs etc) using recommended business intelligence tools. May include guidelines on storing analytics results and other supporting evidence.	Option F	Option F

Question 4

Match each sentence with the associated big data source as either 'raw data' or 'dataset'.

Table 4 - Quiz Q4: Relationship between raw data and datasets

Features and formats	Raw data? Or Dataset?
Also known as cooked data.	Dataset
It does not have any meaning.	Raw data
It is also known as primary data.	Raw data
Manageable and ready for further analysis.	Dataset
Analytical models can be easily run on this data.	Dataset
It is incompatible with conventional data models.	Raw data
Data in its natural, unprocessed and unstructured form.	Raw data
Data that is transformed, cleaned and organised in some way.	Dataset

Question 5

Match each data type example with the correct classification categories of analytics from the list of options provided.

Table 5 Quiz Q5: Classification categories of analytics

Data type example	Answer	Classification categories of analytics
Transcripts of call centre communications with customers	[2] Text	1. Web
Voice recordings of call centre communications with customers	[3] Audio	2. Text
Data from a storefront security camera	[4] Video	3. Audio
Location data from wearable devices	[5]	4. Video

Data type example	Answer	Classification categories of analytics
	Network	
RSS feeds	[1] Web	5. Network

Question 6

Match each description with the correct statistical concepts for big data analysis from the list of options provided.

Table 6 Quiz Q6: Statistical concepts

Description	Answer	Statistical concepts
The middle number in a sorted list of the data.	[3] Median	1. Mode
Sum of all values divided by the number of values.	[7] Arithmetic mean	2. Outlier
The most commonly occurring category or value in a dataset.	[1] Mode	3. Median
This refers to any value that is very distant from the other values in the dataset	[2] Outlier	4. Z score
This is the square root of the variance and is a measure of spread.	[10] Standard deviation	5. Kurtosis
This divides the smallest 25% of the values from the other 75% that are larger. This is commonly known as the 25th percentile.	[8] First quartile	6. Skewness
This divides the smallest 75% of the values from the largest 25%, This is commonly known as the 75th percentile.	[9] Third quartile	7. Arithmetic Mean
A measure used to describe the shape of the distribution in terms of whether it is heavy-tailed (with more outliers) or light-tailed (with fewer outliers).	[5] Kurtosis	8. First Quartile
A standardised value that provides a relative measure of the distance an observation is from the mean, which is independent of the unit of measurement.	[4] Z score	9. Third Quartile
Measures the relative strength of a linear relationship between two numerical variables. A statistical measure of the strength of the relationship between the relative movements of two variables. The possible values range between -1 and 1.	[11] Correlation coefficient	10. Standard Deviation
A measure used to describe the degree of asymmetry observed in a set of data. This value can vary between positive or negative. If this value is '0' that means that the dataset is normally distributed (bell curve).	[6] Skewness	11. Correlation coefficient

Question 7

Match each mathematical equation with the correct statistical concepts used when analysing big data from the list of options (A to H) provided where:

X_i = i^{th} value of the variable X

N = number of values in the population

$\sum_{i=1}^{N} X_i$ = summation of all X_i values in the population.

Statistical concepts

A. Range
B. Z Score
C. First Quartile (Q1)
D. Third Quartile (Q3)
E. Population mean (μ)
F. Interquartile range (IQR)
G. Pearson's Median Skewness
H. Population standard deviation (σ)

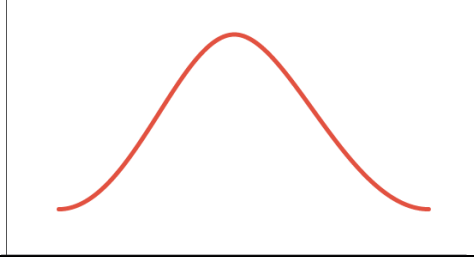
Table 7 Quiz Q7: Statistical concepts

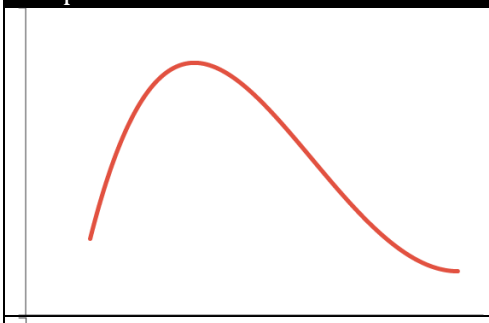
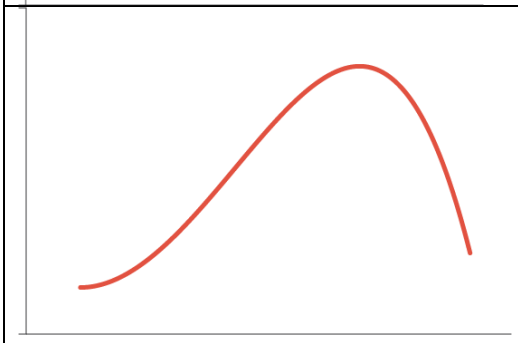
Mathematical equation	Answer	Statistical concepts
$\frac{(N+1)}{4}$	Option C	Option A
$\frac{(X_i - \mu)}{\sigma}$	Option B	Option B
$Q3 - Q1$	Option F	Option C
$\frac{3(N+1)}{4}$	Option D	Option D
$X_{\text{Largest}} - X_{\text{Smallest}}$	Option A	Option E
$\sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$	Option H	Option F
$\frac{\sum_{i=1}^{N} X_i}{N}$	Option E	Option G
$\frac{3(Mean - Median)}{\sigma}$	Option G	Option H

Question 8

Match each graph with the correct statistical concepts relating to big data analytics that indicates the shape of the distribution in a dataset.

Table 8 - Quiz Q8: Statistical concepts

Graph	Answer	Shape of the distribution
	[3] Normally distributed	1. Positively skewed

Graph	Answer	Shape of the distribution
	[1] Positively skewed	2. Negatively skewed
	[2] Negatively skewed	3. Normally distributed

Question 9

Match the listed characteristics with the correct data models (either 'Star schema', 'Snowflake schema' or 'both') commonly used for analysing big data.

Table 9 Quiz Q9: Common models used to analyse big data

Characteristics of the data models	Answer	**Data analysis models**
The is a simple schema design	[1] Star schema	1. Star schema
This is a complex schema design	[2] Snowflake schema	2. Snowflake schema
One fact table is surrounded by multiple dimension tables.	[1] Star schema	3. Both
Requires the tables in the data model to be classified as either *dimension* or *fact*.	[3] Both	
Queries may run slow due to the complex nature of the model.	[2] Snowflake schema	
Queries can run faster in this data model due to its de-normalised data structure.	[1] Star schema	
In this model, only a single join is used to define a relationship between a fact table and a dimension table.	[1] Star schema	
One fact table is surrounded by dimension tables and dimension tables are also surrounded by other dimension tables.	[2] Snowflake schema	

Question 10

Match each description with the correct names of the tools, functions and features of Microsoft Excel software that can be used to perform advanced analytics of external big data, from the options provided.

Table 10 Quiz Q10: Common tools, functions and features of Excel for advanced analytics

Description	Answer	Data analysis tools, functions and features
A function used to compute z-scores easily on a spreadsheet.	[8] STANDARDIZE	1. SKEW.P
A powerful tool to calculate, summarise and analyse data that lets you see comparisons, patterns and trends in the data.	[3] PivotTable	2. Trendline
A function used to return a value that helps to describe the shape of a distribution based on the population of the dataset.	[1] SKEW.P	3. PivotTable
An analytical tool used to draw a line of best fit that represents the behaviours of a set of data to determine if there is a certain pattern.	[2] Trendline	4. PivotChart
A tool that provides graphical representations of the data in associated PivotTables and facilitates comparisons, patterns and trends.	[4] PivotChart	5. Power Pivot
A feature that empowers one to understand data through natural language queries and helps to ask questions about the data without having to write complicated formulas. It provides high-level visual summaries, trends and patterns.	[7] Analyze data	6. Power Query
This is a powerful transformation tool that has the features and capabilities to transform and prepare data for business intelligence projects. It also has the ability to connect to data from a variety of big data sources.	[6] Power Query	7. Analyze data
A **tool that** enables you to take massive volumes of data from different places, analyse the information quickly, and then process it and offer insights. Some features of this tool that enhance data analysis capabilities are; • automatic relationship detection • automatic time grouping • inline creation of DAX calculated measures.	[5] Power Pivot	8. STANDARDIZE

Question 11

Match each description on performing big data analytics with the associated technology and automation tools.

Table 11 - Quiz Q11: Technology and automation tools

Description	Answer	Technology and automation tools
Power BI provides built-in features that use this technology to automate processes and predict outcomes and perform tasks like forms processing, object detection, text, binary classification, anomaly detection and generation of reports on key influencers and decomposition trees.	[3] Artificial Intelligence	1. Cloud technology
Platforms such as Azure, AWS, IBM, Snowflake, Google etc. provides the necessary in-built tools and functionality for creating data models that can scale with the increasing volume of big data.	[1] Cloud technology	2. Machine Learning
Microsoft Power BI built-in tools that leverage this technology enable using the entities defined in Power BI dataflows to train and validate models based on historical	[2] Machine Learning	3. Artificial Intelligence

Description	Answer	Technology and automation tools
datasets.		

Question 12

Match the sources of uncertainty within big data with the correct characteristics of big data (Volume, Variety, Velocity and Veracity).

Table 12 - Quiz Q12: Sources of uncertainty

Sources of uncertainty within big data	Answer	Characteristics of big data
Use of data analysis techniques that are not designed for large-scale databases for scanning and understanding the data at scale.	[2] Volume	1. Variety
Performing conversions between different types of data (e.g. structured, unstructured, semi-structured) using algorithms that have limitations in handling multiple formats.	[1] Variety	2. Volume
Delays in capturing data from systems that process data at different speeds (batch processing systems, real-time systems and near real-time systems)	[4] Velocity	3. Veracity
Data obtained from various sources that are noisy, ambiguous, incomplete or of poor quality where data accuracy and trust is hard to establish.	[3] Veracity	4. Velocity

Question 13

Which of the following can cause uncertainties within big data? Select six (6) correct answers.

a. noisy data
b. clean data
c. correct data
d. incorrect data
e. complete data
f. ambiguous data
g. incomplete data
h. poor quality data
i. inconsistent data
j. good quality data

Question 14

Indicate the correct order of the stages of big data analysis.

Table 13 - Quiz Q14: Stages of big data analysis

Stages of big data analysis	Order
Data communication	5

Stages of big data analysis	Order
Exploratory data analysis	4
Data requirement gathering	1
Data cleaning and processing	3
Data source identification and management	2

Refer to the industry protocols and procedures followed when using **M Query statements** and **Data Analysis Expressions (DAX)** to write queries and scripts for big data testing. Based on this information, answer **questions 15 – 20.**

Question 15

Match each description of the industry protocols (rules) with the metacharacter symbols associated with operators and expressions used when writing basic queries to search combined big data.

Table 14 - Quiz Q15: Industry protocols for writing basic queries

Description of the industry protocols	Answer	Metacharacter symbols *(Operators and wildcards)*
Represents a logic operator for 'OR'	||	1. >
Represents a logic operator for 'AND'	&&	2. /
An arithmetic operator used to perform divisions	/	3. ||
A comparison operator that indicates 'not equal to'	<>	4. ( )
A comparison operator that indicates 'greater than'	>	5. &&
A comparison operator that indicates 'less than or equal to'	<=	6. <>
A type of operator used for precedence order and grouping of arguments	( )	7. <=

Question 16

Indicate whether the following statements are true or false.

Table 15 - Quiz Q16: Industry protocols for writing basic queries

Statement	True or False
In a DAX formula all object names are case-sensitive.	False
Leading or trailing spaces that are not enclosed by name delimiters, brackets or single apostrophes are valid in the names of tables, columns, and measures.	False

Statement	True or False
The following characters are not valid in the names of objects: .,;':/*|?&%$!+=()[]{}<>	True
A common way to add a single line comment into a DAX script is to begin the comment with the // symbol.	True
Adding comments affects the performance of the queries and scripts.	False
A common way to comment out code in a script is to place the code in between /* and */ symbols.	True

Question 17

Match the descriptions with the correct type of join operations when performing merge queries.

Table 16 - Quiz Q17: Industry protocols for writing basic queries

Description	Answer	Type of join operation
Keeps all the rows from the primary table and brings in any matching rows from the related table.	[4] Left outer join	1. Inner join
Keeps all the rows from the related table and brings in any matching rows from the primary table.	[5] Right outer join	2. Fuzzy merge
Brings in all the rows from both the primary and related tables.	[3] Full outer join	3. Full outer join
Brings in only matching rows from both the primary and related tables.	[1] Inner join	4. Left outer join
A smart data preparation feature that uses an algorithm when comparing columns to find matches across the tables that are being merged. This is only supported for merging columns that contain text.	[2] Fuzzy merge	5. Right outer join

Refer to the following tables and answer the following questions **18 to 20.**

Table 17 - Order Detail

Country	State	Cost	Revenue
AUS	NSW	$ 2,000	$ 3,500
AUS	QLD	$ 3,000	$ 5,000
AUS	VIC	$ 1,500	$ 2,500
AUS	ACT	$ 500	$ 1,500
USA	OHI	$ 2,500	$ 4,550
USA	TEX	$ 500	$ 1,750
USA	UTA	$ 3,500	$ 4,550

Question 18

Select the correct DAX function that counts the number of different values in the 'Country' column in the table 'Order Detail'.

Table 18 - Quiz Q18: Industry protocols for writing basic queries

DAX function statement	Answer
= **DISTINCTCOUNT**('Order Detail', [Country])	
= **DISTINCT** (Order Detail[Country])	
= **DISTINCTCOUNT**([Country], 'Order Detail')	
= **DISTINCT**([Country], Order Detail)	
= **DISTINCTCOUNT**('Order Detail'[Country])	X
= **DISTINCT** ('Order Detail', [Country])	

Question 19

Select the correct DAX function that can be used to calculate the 'Gross Profit' from all the store locations.

Note: Gross Profit = Total Revenue – Total Cost

Table 19 - Quiz Q19: Industry protocols for writing basic queries

DAX function	Answer
= **ADD**('Order Detail'[Revenue - Cost])	
= **SUM**('Order Detail'[Revenue - Cost])	
= **ADD**(Order Detail[Revenue]) – **ADD** (Order Detail[Cost])	
= **CALCULATE**('Order Detail'[Revenue] – 'Order Detail'[Cost])	
= **SUM** ('Order Detail'[Revenue]) – **SUM** ('Order Detail'[Cost])	X

Question 20

Select the correct DAX function that calculates the 'Gross Profit Margin'. If either of the values in the function is a '0' then ensure that the calculation would return '0' as the result.

Note: Assume that you have previously created two calculated measures called 'Gross Profit' and 'Total Sales'. The equation used to calculate 'Gross Profit Margin' is as follows.

$$Gross\ Profit\ Margin = \frac{[Gross\ Profit]}{[Total\ Sales]}$$

Table 20 - Quiz Q20: Industry protocols for writing basic queries

DAX function statement	Answer
= **DIVIDE** ([Gross Profit], [Total Sales])	
= **DIVIDE** ([Total Sales], [Gross Profit])	
= **DIVIDE** ([Gross Profit], [Total Sales], 0)	X
= **DIVIDE** ('Profit Summary'[Total Sales], 'Profit Summary'[Gross Profit])	
= **DIVIDE** ('Profit Summary'[Gross Profit], 'Profit Summary'[Total Sales], 0)	

Assessment checklist:

Students must have completed all questions within this assessment before submitting. This includes:

1	20 quiz questions to be completed.	☐

Congratulations you have reached the end of Assessment [1]!