EXTERNAL INTEGRATED SUMMATIVE ASSESSMENT (EISA)

118708 - OCCUPATIONAL CERTIFICATE: DATA SCIENCE PRACTITIONER, NQF LEVEL 5, CREDITS 185

DATA SCIENCE PRACTITIONER EXEMPLAR ASSESSMENT

Duration: 3 Hours

Total Marks: 100

Pass Mark: 60

CANDIDATE INFORMATION													
SURNAME													
NAMES													
ID NUMBER													
EISA REGISTRATION NUMBER													
ASSESSMENT CENTRE													
ASSESSMENT CENTRE ACCREDITATION NUMBER													

INSTRUCTIONS TO CANDIDATES

1. Candidates must arrive at the assessment centre at least 30 minutes before the scheduled start time.
2. Candidates must present a valid identification document (ID, passport, or driver's licence) for verification.
3. Candidates must sign the attendance register before entering the assessment venue.
4. Candidates must follow all instructions provided by the invigilator.
5. Candidates must take the seat allocated to them by the invigilator.
6. Candidates may not change seats without permission from the invigilator.
7. Candidates must log into the e-assessment platform using the credentials provided.
8. Candidates must not share login details with any other person.
9. Candidates must ensure that only the authorised assessment platform is open on the computer.
10. Candidates may not open additional browser tabs, applications, or websites during the assessment.
11. Candidates may not use mobile phones, smart watches, tablets, or any other electronic devices during the assessment.
12. Candidates must switch off and store mobile phones in the area designated by the invigilator.
13. Candidates may not bring notes, textbooks, bags, or unauthorised materials into the assessment venue.
14. Candidates may not communicate with other candidates during the assessment.
15. Candidates may not attempt to copy from another candidate's screen.
16. Candidates may not receive assistance from any person during the assessment.
17. Candidates must not attempt to access external information sources, including the internet or artificial intelligence tools.

18. Candidates must raise their hand to request assistance from the invigilator.
19. Candidates may not leave the assessment venue without permission from the invigilator.
20. Candidates who leave the venue during the assessment may not take any assessment material with them.
21. Candidates must submit the assessment electronically before the allocated time expires.
22. Candidates must remain seated until instructed by the invigilator to leave the venue.
23. Candidates may not copy, photograph, screenshot, or distribute assessment content.
24. Any form of cheating, impersonation, or misconduct will result in disqualification from the assessment.
25. Candidates must comply with all assessment policies of the Assessment Quality Partner (AQP) appointed under the Quality Council for Trades and Occupations.

Question 1 **[33]**

SCENARIO

A digital health company has developed a smartwatch that continuously records daily health metrics such as heart rate, step count, sleep duration, calories burned, and activity levels. The data collected from these devices is stored in a central system for further health monitoring and analysis.

You have been appointed as a Data Science Practitioner to analyze data generated by these smartwatch devices. The dataset has been provided in a file named smartwatch_health_data.csv. The file contains raw data collected from multiple users and sources and requires preparation before meaningful analysis can be performed.

Question 1.1 **[8]**

1.1.1 Identify two problems that may occur if management expectations are not clearly defined before data analysis begins (2)

.

1.1.2 Explain two reasons why it is important to set well-defined goals before analyzing health-related data collected from smart devices. (4)

1.1.3 Formulate one well-defined goal for the analysis of the smartwatch health data. Your goal must be clear and measurable (2)

.

1.2 **[8]**

1.2.1 Open the smartwatch_health_data.csv file and identify the column that best represents the primary data collected directly from the smartwatch sensors. Classify this column's data source type and justify your answer. (2)

1.2.2 Using the smartwatch_health_data.csv dataset, identify two columns whose data could be combined with data from external sources to improve the health monitoring analysis. For each column, name a suitable external source and explain what additional value it would bring to the analysis (4)

.

1.2.3 Write Python code to load the smartwatch_health_data.csv file into your working environment. Based on your initial inspection of the dataset, recommend whether the data is ready for analysis or not, and provide evidence from your output to support your recommendation. (2)

1.3 **[8]**

1.3.1 Run two different Python operations on the smartwatch_health_data.csv dataset and for each operation, describe what it reveals about the structure or content of the data. Include your code and the output in your answer (2)

1.3.2 Using the dataset loaded in Question 1.2, write Python code to add a new column called heart_rate_status that classifies each record as either 'Normal' or 'Abnormal' based on whether the heart rate is within the normal resting range of 60 to 100 beats per minute (bpm). Display the first ten records of the updated dataset. (4)

1.3.3 The data science team requires a summarized view of the dataset before proceeding with analysis. Write Python code to produce a statistical summary of the smartwatch_health_data.csv dataset and evaluate whether the summary reveals any concerns about the data. Justify your answer (2)

1.4 **[9]**

1.4.1 Open the smartwatch_health_data.csv file and examine its contents. Identify THREE specific data quality issues you can observe in the dataset and for each issue, explain how it could negatively affect the outcome of the health data analysis. (3)

1.4.2 Using the dataset loaded in Question 1.2, write Python code to clean and prepare the data by addressing any missing values, duplicate records, and inconsistent data formats you find. Your code must display the shape of the dataset before and after cleaning to show the impact of your changes (5)

.

1.4.3 Justify one cleaning decision you made in Question 1.4.2 and explain why the method you chose was the most appropriate for this dataset. (1)

Question 2 **[33]**

Scenario

FinSight Analytics is a South African financial technology company that helps individuals track and improve their personal financial health. The company collects monthly financial snapshots from its 3 000 registered users, capturing data such as income, expenses, savings behavior, credit profiles, and spending categories across different economic scenarios including normal, inflation, and recession periods.

The data science team has been asked to analyze the user financial data to uncover patterns and trends that could help the company make informed business decisions. The dataset has been provided in a file named personal_finance_tracker_dataset.csv.

Question 2.1 **[8]**

2.1.1 Examine the personal_finance_tracker_dataset.csv file. Identify TWO columns that would be most useful for querying user financial behavior and explain why each column is valuable for this purpose (2)

2.1.2 Load the personal_finance_tracker_dataset.csv file into a SQL database. Write a SQL query to retrieve all records where the financial_scenario is 'recession' and the fraud_flag is equal to 1. Your query must display the user_id, date, monthly_income, onthly_expense_total and fraud_flag columns only. Order the results by monthly_income in descending order. (4)

2.1.3 Write a SQL query to retrieve the total number of users for each financial_scenario in the dataset. Your query must display the financial_scenario and the total count, ordered from the highest to the lowest count. (2)

Question 2.2 **[8]**

2.2.1 Open the personal_finance_tracker_dataset.csv file and examine the financial_stress_level and savings_rate columns. Describe the general tendency you observe between these two columns and explain what this could mean for FinSight Analytics users (2)

2.2.2 Using SQL, calculate the average monthly_income, monthly_expense_total and actual_savings for each financial_scenario in the dataset. Display the results in a structured format and identify which financial scenario shows the most concerning tendency for users. Justify your answer (4)

.

2.2.3 Using SQL, determine which income_type has the highest average debt_to_income_ratio across all financial scenarios. Based on your output, explain what this tendency suggests about the financial behavior of users with that income type. (2)

Question 2.3 **[8]**

2.3.1 Open the personal_finance_tracker_dataset.csv file and examine the credit_score and debt_to_income_ratio columns. Describe one trend you observe in these columns and explain whether this trend could be used for diagnostic or predictive analytics. Justify your answer (2)

2.3.2 Using either SQL or Python, analyze the personal_finance_tracker_dataset.csv dataset to identify the trend in average credit_score across different financial_scenario values over time. Display your results in a structured format and explain what the trend suggests about the relationship between economic conditions and user creditworthiness (4)

.

2.3.3 Using either SQL or Python, calculate the average financial_stress_level for each financial_scenario in the dataset. Based on your output, identify which scenario poses the greatest diagnostic concern for FinSight Analytics and explain what informed business decision the company could make based on this finding. (2)

Question 2.4 **[9]**

2.4.1 The FinSight Analytics data science team needs to verify that the analysis performed in Questions 2.2 and 2.3 produces accurate and reliable results. Identify two ways in which the accuracy of a data analysis output can be tested before presenting findings to a non-technical audience (2)

.

2.4.2 Using Python, create TWO different visualizations based on the Personal_finance_tracker_dataset.csv dataset that clearly communicate patterns or trends in user financial behavior to a non-technical audience at FinSight Analytics. For each visualization, include a title, labelled axes, and a brief written explanation of what the visual reveals about the data. (5)

2.4.3 Based on the patterns and trends uncovered in your analysis across Questions 2.1 to 2.3, write a short insight report of no more than five sentences that communicates your key findings to the FinSight Analytics management team. Your report must be written for a non-technical audience and must include at least one recommendation for the business (2)

Question 3 **[34]**

Scenario

AutoInsight is a global automotive market research firm that provides strategic intelligence to vehicle manufacturers, investors, and policy makers. The firm specializes in analyzing sales trends, electric vehicle adoption, and macroeconomic influences across major automotive markets worldwide.

The data science team has been asked to prepare a descriptive analytics report on BMW Group's global sales performance from 2018 to 2025. The report must communicate key patterns and trends through visualizations and data storytelling to both technical and non-technical stakeholders. The dataset has been provided in a file named bmw_global_sales_2018_2025.csv.

Question 3.1 **[4]**

3.1.1 Explain what it means for a visual element to accurately reflect the outcomes of a data analysis

(1)

3.1.2 Open the bmw_global_sales_2018_2025.csv file and examine its contents. A colleague has produced the following summary statement: 'The X5 is BMW's best-selling model globally across all regions and all years in the dataset.' Using any suitable platform, write code or apply a method to verify whether this statement is accurate. Present your output as evidence and explain whether the visual or analytical result confirms or contradicts the statement. (3)

Question 3.2 **[6]**

3.2.1 Identify ONE type of chart that would be most appropriate for displaying BMW's total Units_Sold per Region over time and explain why this chart type is suitable for this purpose. (1)

3.2.2 Using the bmw_global_sales_2018_2025.csv dataset, create a pivot table that summarizes the total Revenue_EUR per Region for each Year. Your pivot table must be clearly formatted and must allow the data to be filtered by Model. (3)

3.2.3 Using the pivot table created in Question 3.2.2, create ONE visualization that clearly communicates the trend in BMW's total revenue across regions from 2018 to 2025. Your visualization must include a title, labelled axes, and a legend. Explain what business concept the visualization communicates to a non-technical audience at AutoInsight. (2)

Question 3.3 **[6]**

3.3.1 At the beginning of the data analysis cycle, AutoInsight defined the following key question: 'Which BMW model generates the highest revenue across all regions and how has this changed over time?' Explain what criteria should be used to determine whether a data analysis report successfully answers this question (1)

.

3.3.2 Using the bmw_global_sales_2018_2025.csv dataset, write code or apply a method to determine which BMW Model generated the highest total Revenue_EUR across all regions for each Year from 2018 to 2025. Display your results in a structured format and explain whether your output successfully answers the key question defined in Question 3.3.1. (3)

3.3.3 Based on your output from Question 3.3.2, evaluate whether the analysis fully addresses the key question defined at the beginning of the data analysis cycle. Identify ONE limitation of your analysis and explain how it could be addressed to improve the completeness of the report. (2)

Question 3.4 **[6]**

3.4.1 Distinguish between diagnostic analysis and predictive analysis. For each, provide one example of how it could be applied to the bmw_global_sales_2018_2025.csv dataset to support AutoInsight's reporting objectives. (2)

3.4.2 Using the bmw_global_sales_2018_2025.csv dataset, write code or apply a method to analyze the trend in BEV_Share across all regions from 2018 to 2025. Display your results in a structured format and use the trend you identify to predict whether BEV penetration is likely to increase or decrease beyond 2025. Support your prediction with statistical evidence from your output. (3)

3.4.3 Based on the trend identified in Question 3.4.2, identify ONE external factor present in the dataset that could influence the accuracy of your prediction and explain how it could affect the outcome (1)

Question 3.5 **[6]**

3.5.1 Explain what data storytelling is and state one reason why it is important when presenting data analysis findings to a non-technical audience such as AutoInsight's clients. (1)

3.5.2 Using the bmw_global_sales_2018_2025.csv dataset, identify ONE significant pattern or trend in the data that would be valuable to AutoInsight's clients. Create a visualization that clearly communicates this pattern and write a short data story of no more than three sentences that explains the insight to a non-technical audience. Your data story must reference specific values from your output. (3)

3.5.3 AutoInsight's management team has requested a one-paragraph executive summary of the BMW global sales analysis. Using the patterns and trends uncovered across Questions 3.1 to 3.4, write an executive summary that communicates the most important findings in plain language. Your summary must include at least one data-driven recommendation for a BMW market strategy decision (2)

Question 3.6 **[6]**

3.6.1 Open the bmw_global_sales_2018_2025.csv file and examine the BEV_Share, GDP_Growth and Fuel_Price_Index columns. Identify TWO potential process improvement opportunities for AutoInsight based on patterns you observe in these columns. For each opportunity, explain how it could improve the company's analytical or reporting processes (2)

3.6.2 Using the bmw_global_sales_2018_2025.csv dataset, write code or apply a method to identify which Region has shown the slowest growth in BEV_Share between 2018 and 2025. Based on your output, propose ONE specific process improvement that AutoInsight could recommend to BMW to accelerate EV adoption in that region. Justify your recommendation with evidence from your output. (3)

3.6.3 The AutoInsight data science team has completed their analysis of BMW's global sales data. Identify ONE new data source that could be integrated into future analyses to improve the quality and depth of AutoInsight's reporting. Explain what additional insights this data source would enable. (1)

Total Marks 100